

AD-A164 346

A STABLE AND EFFICIENT ALGORITHM FOR NONLINEAR
ORTHOGONAL DISTANCE REGRES. (U) COLORADO UNIV AT
BOULDER DEPT OF COMPUTER SCIENCE P T BOGGS ET AL.

1/1

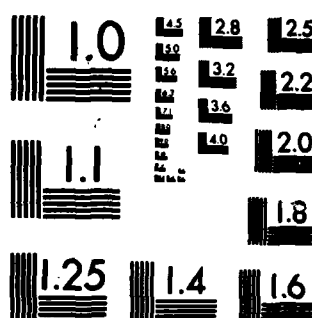
UNCLASSIFIED

DEC 85 CU-CS-317-85 ARO-18197. 8-MA

F/G 12/1

NL

						END							
						FILED							
						DTN							



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AD-A164 346

UNIVERSITY OF COLORADO

A STABLE AND EFFICIENT ALGORITHM FOR NONLINEAR ORTHOGONAL DISTANCE REGRESSION

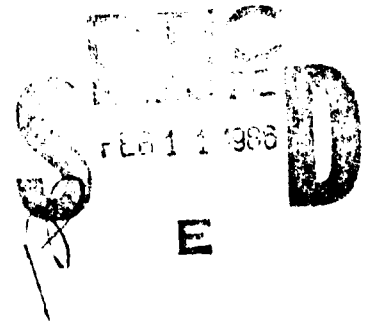
Paul T. Boggs *
Richard H. Byrd **
Robert B. Schnabel ***

CU-CS-317-85 December, 1985

DEPARTMENT OF COMPUTER SCIENCE
100 S. HONOLULU
UNIVERSITY OF COLORADO, BOULDER
BOULDER, COLORADO 80509

Technical Report

FILE COPY



86 2 7 131

A STABLE AND EFFICIENT ALGORITHM
FOR NONLINEAR ORTHOGONAL
DISTANCE REGRESSION

Paul T. Boggs *
Richard H. Byrd **
Robert B. Schnabel ***

CU-CS-317-85 December, 1985

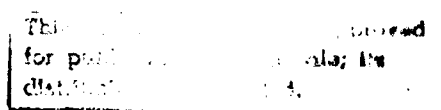
DTIC
ELECTE
FEB 11 1986
S D
E

* Scientific Computing Division, National Bureau of Standards, Gaithersburg, Maryland 20899

** Department of Computer Science, University of Colorado, Boulder, Colorado 80309

*** Department of Computer Science, University of Colorado, Boulder, Colorado 80309 and
Scientific Computing Division, National Bureau of Standards, Boulder, Colorado 80309

Research supported by NSF grant DCR-8403483 and ARO contract DAAG 29-81-K0108



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD-A164346

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO 18197.8-MA	2. GOVT ACCESSION NO. N/A	3. RECIPIENT'S CATALOG NUMBER N/A
4. TITLE (and Subtitle) A Stable And Efficient Algorithm For Nonlinear Orthogonal Distance Regression		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG REPORT NUMBER
7. AUTHOR(s) Paul T. Boggs Richard H. Byrd Robert B. Schnabel		8. CONTRACT OR GRANT NUMBER(s) DAAG 29-81-K-0108
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Colorado Dept. of Computer Science Campus Box 430 Boulder, CO 80309		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE December 1985
		13. NUMBER OF PAGES 40
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report) NA		
18. SUPPLEMENTARY NOTES The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Data Fitting, Orthogonal, Regression, Trust Region		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) One of the most widely used methodologies in scientific and engineering research is the fitting of equations to data by least squares. In cases where significant observation errors exist in all data (independent) variables, however, the ordinary least squares approach, where all errors are attributed to the observation (dependent) variable, is often inappropriate. An alternate approach, suggested by several researchers, involves minimizing the sum of squared orthogonal distances between each data point and the curve described by		

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT CONTINUED


the model equation. We refer to this as orthogonal distance regression (ODR). This paper describes a method for solving the orthogonal distance regression problem that is a direct analog of the trust region Levenberg-Marquardt algorithm. The number of unknowns involved is the number of model parameters plus the number of data points, often a very large number. By exploiting sparsity, however, our algorithm has a computational effort per step which is of the same order as required for the Levenberg-Marquardt method for ordinary least squares. We prove our algorithm to be globally and locally convergent, and perform computational tests that illustrate some differences between the two approaches.

Accession For	
NTIS	<input checked="" type="checkbox"/>
DTIC	<input type="checkbox"/>
Other	<input type="checkbox"/>
JPRS	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

(A)

Abstract

One of the most widely used methodologies in scientific and engineering research is the fitting of equations to data by least squares. In cases where significant observation errors exist in all data (independent) variables, however, the ordinary least squares approach, where all errors are attributed to the observation (dependent) variable, is often inappropriate. An alternate approach, suggested by several researchers, involves minimizing the sum of squared orthogonal distances between each data point and the curve described by the model equation. We refer to this as orthogonal distance regression (ODR). This paper describes a method for solving the orthogonal distance regression problem that is a direct analog of the trust region Levenberg-Marquardt algorithm. The number of unknowns involved is the number of model parameters plus the number of data points, often a very large number. By exploiting sparsity, however, our algorithm has a computational effort per step which is of the same order as required for the Levenberg-Marquardt method for ordinary least squares. We prove our algorithm to be globally and locally convergent, and perform computational tests that illustrate some differences between the two approaches.



When observations are corrupted
By errors random i.i.d.
Least Squares provides the answer
A true blue friend it be.

But when the independents
In addition come unclean
Nonlinearities enter
And friends desert the scene.

Now in your darkest hour
When hope is all but gone
This paper comes to show you
How estimation can go on.

What is the shining hope
That beacon from afar?
Is it that new found wonder
That's known as ODR?

Yes! That's right! You've got it!
And we tell all below
From problem through solution
With codes to make it go.

So calm your pounding heart
And mop your sweating brow.
Your data can be handled
And we will show you how.

1. Introduction

The problem of fitting a model to data with errors in the observations has a rich history and a considerable literature. The problem where there are also errors in the independent variables at which these observations are made, however, has only relatively recently been given attention. In this paper, we consider a general form of this extended problem and provide an efficient and stable algorithm for its solution. Several names for this extended problem have been suggested; we prefer orthogonal distance regression (ODR).

Errors in independent variables virtually always occur, but are often ignored in order that classical or ordinary (linear or nonlinear) least squares (OLS) techniques can be applied (see, e.g., [LawH74], [Ste73], [Mor77], [DenGW81]). Also, if these errors are small with respect to those in the observed variables, then ignoring them does not usually seriously degrade the accuracy of the estimates. In some fields, however, measurement techniques are sufficiently accurate that errors in the independent variables are not insignificant compared to those in the observations. Examples at the National Bureau of Standards (NBS) include the calibration of electronic devices, flow-meters and calorimeters. Another class of examples comes from curve and surface fitting problems.

We first develop a formal statement of the ODR estimation problem and briefly discuss its application to statistical estimation and to curve fitting. The main contributions of the paper are the derivation and convergence analysis of a highly efficient algorithm for solving ODR problems (Sections 2 and 3). In Section 4, the results of some computations are shown which illustrate the performance of the algorithm and allow some comparisons with ordinary least squares.

Observations in applied science are often thought of as satisfying a mathematical model of the form

$$(1.1) \quad y = f(x, \beta)$$

where y is taken to be the "observed" value, or independent variable; and $\beta \in R^p$ is the set of parameters to be estimated. The function f is not assumed to be linear, but is assumed to be smooth. The data are simply the pairs (x_i, y_i) , $i = 1, \dots, n$. Typically the number of data points, n , is far greater than the number of parameters, p .

In the classical case, only the observations y_i are assumed to be contaminated with errors. If these errors are additive and the mathematical model is exact then

$$(1.2) \quad y_i = f(x_i, \hat{\beta}) + \hat{\epsilon}_i \quad i = 1, \dots, n$$

for some correct value of the parameters $\hat{\beta}$. If in addition the errors are normally distributed with mean 0 and variance $\sigma^2 I$, then maximum likelihood estimate of $\hat{\beta}$ is the solution to the least squares problem

$$(1.3) \quad \min_{\beta} \sum_{i=1}^n [y_i - f(x_i, \beta)]^2.$$

If f is a linear function of $\hat{\beta}$ then this is a classical linear least squares problem, otherwise it is a classical nonlinear least squares problem. Even when the above assumptions on the

model or the errors are not satisfied, problem (1.3) is the most frequently used method for parameter estimation.

In the more general situation, the measurements of the independent variables x_i are also assumed to contain errors. If we assume that y_i has unknown additive error $\hat{\epsilon}_i$ and that x_i has unknown additive error $\hat{\delta}_i$, then (1.2) becomes

$$(1.4) \quad y_i = f(x_i + \hat{\delta}_i; \beta) + \hat{\epsilon}_i.$$

An intuitively reasonable way to select the parameters in this case is to choose the β that causes the sum of the squares of the orthogonal distances from the data points (x_i, y_i) to the curve $f(x, \beta)$ to be minimized. (See Figure 1.) If r_i is the orthogonal distance from (x_i, y_i) to the curve, then

$$r_i^2 = \epsilon_i^2 + \delta_i^2, \quad i = 1, \dots, n$$

where ϵ_i and δ_i solve

$$(1.5) \quad \begin{aligned} & \min_{\epsilon_i, \delta_i} (\epsilon_i^2 + \delta_i^2) \\ & \text{subject to } f(x_i + \delta_i; \beta) + \epsilon_i = y_i. \end{aligned}$$

The constraint in (1.5) ensures that the distance r_i connects the point (x_i, y_i) to the curve. The minimization ensures that r_i is the radius of the smallest circle centered at (x_i, y_i) which is tangent to the curve $f(x_i; \beta)$. (See Figure 2). Therefore, the parameters β that cause the sum of the squares of the orthogonal distances from the data points to the curve to be minimized are found by solving

$$(1.6) \quad \begin{aligned} & \min_{\beta, \epsilon, \delta} \sum_{i=1}^n r_i^2 = \min_{\beta, \epsilon, \delta} \sum_{i=1}^n (\epsilon_i^2 + \delta_i^2) \\ & \text{subject to } y_i = f(x_i + \delta_i; \beta) + \epsilon_i, \quad i = 1, \dots, n. \end{aligned}$$

Since the constraints in (1.6) are simple linear constraints in ϵ_i , we solve for ϵ_i and eliminate both these variables and all of the constraints thereby obtaining

$$\min_{\beta, \delta} \sum_{i=1}^n \left[(f(x_i + \delta_i; \beta) - y_i)^2 + \delta_i^2 \right]$$

which is now an unconstrained minimization problem.

Two slight extensions to this form constitute the ultimate problem to be considered. The first allows the possibility that $x_i \in R^m$ rather than R^1 . Therefore, $\delta_i \in R^m$ and instead of δ_i^2 in (1.7) we have $\delta_i^T \delta_i = \sum_{j=1}^m \delta_{ij}^2$. (The superscript T denotes transpose.) The second extension merely admits a general weighting scheme on the problem. The form we have chosen results in the general nonlinear ODR problem

$$(ODR) \quad \min_{\beta, \delta} \sum_{i=1}^n w_i^2 \left[(f(x_i + \delta_i; \beta) - y_i)^2 + \delta_i^T D_i^2 \delta_i \right]$$

where $w_i > 0, i = 1, \dots, n$ and $D_i = \text{diag}\{d_{ij} > 0, j = 1, \dots, m\}, i = 1, \dots, n$, i.e., D_i is a diagonal matrix of order m . It follows that the vectors $y, w \in R^n$ and $x, \delta \in R^{nm}$ and that $\delta_i^T D_i \delta_i = \sum_{j=1}^m \delta_{ij}^2 d_{ij}^2$.

While we have not assumed that f is linear, it is important to note that (ODR) is a nonlinear optimization problem even if f is the simple linear function

$$y = \beta_1 x + \beta_2$$

since we then have that

$$y_i = \beta_1(x_i + \delta_i) + \beta_2 + \epsilon_i.$$

Clearly the product of β_1 and δ_i is an unavoidable nonlinearity.

ODR problems have been considered by statisticians, usually under the rubric errors in variables. Most of this effort, however, has been devoted to linear models, i.e., when f is linear in β . (See e.g., [Mor71], [KenSO83], [Bar74, p.67] and [Ful86].) As in the classical nonlinear least squares case, little theory on the statistical properties of the solution appears to exist. It is known that if both ϵ and δ are normally distributed with mean zero and variances $\sigma_\epsilon I$ and $\sigma_\delta I$ respectively, then the solution of (ORD) with $w_i = 1$ and $D_i = (\sigma_\epsilon / \sigma_\delta) I, i = 1, \dots, n$ is a maximum likelihood estimate of the parameters. Unfortunately, as in the nonlinear classical case, no generally valid, computationally efficient, inferential statistical tests are known.

Independent of statistical considerations, ODR has potentially significant applications in curve and surface fitting. Consider, for example, the problem of finding the parabola which best fits the given set of points (see Figure 3.). (We have seen this problem arise from a dental application.) Here it is clear that ordinary least squares will unduly weight the top data points, while fitting in the horizontal direction would unduly weight the bottom data points. An orthogonal measure of distance alleviates these problems and provides a reasonable fit. A related case is the problem of fitting near an asymptote as illustrated in Figure 4. Orthogonal distances here prevent the undue influence of points close to the asymptote. This problem is discussed further in Section 4.

The literature contains several algorithms for solving (ODR) and related problems. For example, Golub and Van Loan [GolV83] give a singular value decomposition procedure for the problem when f is linear. They refer to this problem as total least squares. Britt and Luecke [BriL73] consider the nonlinear case as well as the nonlinear implicit case and present an algorithm. Recently, Schwetlick and Tiller [SchT85] proposed an algorithm similar to the one here for the nonlinear problem. Our algorithm, however, does not make use of the singular value decomposition and it does incorporate a full trust region strategy.

2. The Algorithm

In order to solve the minimization problem (ODR),

$$(2.1) \quad \min_{\beta, \delta} \sum_{i=1}^n w_i^2 \left[(f(x_i + \delta_i; \beta) - y_i)^2 + \delta_i^T D_i^2 \delta_i \right]$$

we first express it in a more convenient form and simplify the notation. Next, we give an overview of the iteration which is based on the trust region -Levenberg-Marquardt strategy

popularized by Moré [Mor77]. (See also [Heb73], [DenS83].) We then show how to modify this technique to obtain an algorithm which requires the same order of work per iteration as these algorithms applied to the same problem without allowing changes to x_i . That is, if the δ_i 's are held fixed at zero, ODR reduces to OLS and trust region methods require $O(np^2)$ operations per iteration. Our algorithm, by exploiting the structure of (ODR), still requires only $O(np^2) + O(nm)$ operations per iteration to solve the problem.

While we have designed and implemented the algorithm to handle the full generality of (2.1), the notation is considerably simplified by assuming $x_i \in R^1$. We temporarily make this assumption and rewrite (2.1) into the form of an OLS problem by the following device. Let

$$(2.2) \quad g_i(\beta, \delta) = \begin{cases} w_i (f(x_i + \delta_i; \beta) - y_i), & i=1, \dots, n \\ w_{i-n} d_{i-n} \delta_{i-n}, & i=n+1, \dots, 2n. \end{cases}$$

Also let $G : R^{p+n} \rightarrow R^{2n}$ have component functions $g_i(\eta)$ where $\eta = \begin{pmatrix} \beta \\ \delta \end{pmatrix}$. Now (ODR) becomes

$$(2.3) \quad \min_{\eta} \|G(\eta)\|^2 = \min_{\beta, \delta} \sum_{i=1}^{2n} (g_i(\beta, \delta))^2$$

which is an OLS problem with $(p+n)$ parameters and $2n$ equations. (In all cases in this paper, $\|\cdot\|$ denotes the ℓ_2 vector or matrix norm.) Direct application of trust region methods to (2.3) would require $O(2n(n+p)^2)$ operations per iteration which rapidly becomes prohibitive if n is large. (Recall that n is usually far greater than p in practice.)

The basic idea of a trust region strategy is to choose as the step that vector which minimizes a linear approximation to G over a region in which the linearization is a "reasonable" approximation to G . Specifically, if $G'(\eta^c) \in R^{2n \times (n+p)}$ is the Jacobian matrix of G evaluated at the current iterate, η^c , then the step z is chosen by solving

$$(2.4) \quad \begin{aligned} & \min_z \|G(\eta^c) + G'(\eta^c)z\|^2 \\ & \text{subject to } \|Zz\| \leq \tau \end{aligned}$$

where Z is a nonsingular (usually diagonal) scaling matrix and τ is the trust region radius. It is easy to show that the solution to (2.4) is given by the $z(\alpha)$ satisfying

$$(2.5) \quad (G'(\eta^c)^T G'(\eta^c) + \alpha Z^T Z) z(\alpha) = -G'(\eta^c)^T G(\eta^c)$$

where $\alpha > 0$ is the Lagrange multiplier for the inequality constraint. Note that if $\|z(0)\| \leq \tau$, $\alpha = 0$ and the constraint is inactive. Otherwise $\alpha > 0$ and the constraint is active. Equation (2.5) is the famous Levenberg-Marquardt formula, but this derivation has given rise to more stable and robust implementations. (See, e.g., [Mor77] and [DenS83]). Clearly (2.5) can be regarded as the "normal equations" for the extended least squares problem,

$$(2.6) \quad \begin{bmatrix} G'(\eta^c) \\ \alpha Z \end{bmatrix} z = - \begin{bmatrix} G \\ 0 \end{bmatrix}$$

where " \approx_2 " means "equal in the least squares sense."

Our implementation is based on the careful exploitation of the structure of the extended Jacobian matrix in (2.6). From (2.2) we have that

$$G'(\eta^c) = \begin{pmatrix} J & V \\ 0 & D \end{pmatrix}$$

where

$$J \in R^{n \times p} : J_{ij} = \frac{\partial g_i(\beta, \delta)}{\partial \beta_j} = \frac{\partial f(x_i + \delta_i; \beta)}{\partial \beta_j},$$

$$i = 1, \dots, n, \quad j = 1, \dots, p;$$

$$V \in R^{n \times n} : V_{ij} = \frac{\partial g_i(\beta, \delta)}{\partial \delta_j} = \frac{\partial f(x_i + \delta_i; \beta)}{\partial \delta_j},$$

$$i = 1, \dots, n \quad j = 1, \dots, n;$$

$$D \in R_{n \times n} : D = \text{diag}\{w_i d_i, i = 1, \dots, n\}.$$

Here, we have omitted the arguments of J and V for the sake of clarity. Observe that since g_i only depends on δ_i , $i = 1, \dots, n$,

$$V = \text{diag} \left\{ \frac{\partial g_i(\beta, \delta)}{\partial \delta_i}, i = 1, \dots, n \right\}.$$

The structure of $G'(\eta^c)$, therefore, is

$$G'(\eta^c) = \begin{bmatrix} \boxed{J} & \begin{array}{c} \diagup V \diagdown \\ \diagdown D \diagup \end{array} \end{bmatrix}.$$

Commensurate with this partitioning of $G'(\eta^c)$, η^c is naturally partitioned into components $(\beta^c, \delta^c)^T$ and the step z into a step in β , say s , and a step in δ , say t . Furthermore, we allow for s to be scaled by a nonsingular diagonal scaling matrix S and t by a nonsingular diagonal matrix T . Thus (2.6) becomes

$$(2.7) \quad \begin{bmatrix} J & V \\ 0 & D \\ \alpha S & 0 \\ 0 & \alpha T \end{bmatrix} \begin{bmatrix} s \\ t \end{bmatrix} \approx_2 - \begin{bmatrix} G_1 \\ G_2 \\ 0 \\ 0 \end{bmatrix}$$

where G_1 is the first n components of G and G_2 is the last n components.

Now, if $x_i \in R^m$, then (2.7) will have the same form except that $V \in R^{n \times nm}$; $T, D \in R^{nm \times nm}$ are still diagonal; and V , instead of being diagonal, has the "staircase" structure which is illustrated for $n = 4$ and $m = 3$ as follows:

$$V = \begin{bmatrix} xxx & & & \\ & xxx & & \\ & & xxx & \\ & & & xxx \end{bmatrix} = \begin{bmatrix} V_1 & & & \\ & V_2 & & \\ & & V_3 & \\ & & & V_4 \end{bmatrix}$$

where $V_i \in R^{1 \times m}$. The rest of the development now allows $x_i \in R^m$.

To derive an efficient procedure for solving (2.7), we first form the normal equations associated with (2.7):

$$(2.8) \quad \begin{pmatrix} J^T J + \alpha S^2 & J^T V \\ V^T J & V^T V + D^2 + \alpha T^2 \end{pmatrix} \begin{pmatrix} s \\ t \end{pmatrix} = - \begin{pmatrix} J^T G_1 \\ V^T G_1 + D G_2 \end{pmatrix}.$$

Let $P = V^T V + D^2 + \alpha T^2$. Solving the bottom equations of (2.8) for t in terms of s yields

$$(2.9) \quad t = -P^{-1}(V^T G_1 + D G_2 + V^T J s)$$

and then substituting (2.9) into the top equations of (2.8) gives

$$(2.10) \quad [J^T(I - V P^{-1} V^T)J + \alpha S^2] s = -J^T G_1 + J^T V P^{-1}(V^T G_1 + D G_2).$$

Thus s is the solution to the least squares problem.

$$(2.11) \quad \begin{pmatrix} (I - V P^{-1} V^T)^{1/2} J \\ \alpha^{1/2} \end{pmatrix} s = \begin{pmatrix} -J^T G_1 + J^T V P^{-1}(V^T G_1 + D G_2) \\ 0 \end{pmatrix}.$$

The following propositions provide the necessary tools for the accurate and efficient solution of (2.9) and (2.11).

Proposition 2.1. Let $E = D^2 + \alpha T^2$. Then, $V P^{-1} V^T = \text{diag} \left\{ \frac{\omega_i}{1 + \omega_i}, i = 1, \dots, n \right\}$ where

$$\omega_i = \sum_{j=1}^m \frac{V_{ij}^2}{E_{(i-1)m+j}}, \quad i = 1, \dots, n.$$

Proof. We have that E is a nonsingular diagonal matrix and that $V^T V$ is block diagonal where the blocks are $m \times m$ rank 1 matrices. Thus P^{-1} can be calculated by the Sherman-Morrison-Woodbury formula, (see, e.g., [DenS83, page 188] or [OrtR70, page 50]) as

$$(2.12) \quad \begin{aligned} P^{-1} &= (V^T V + E)^{-1} \\ &= E^{-1} - E^{-1} V^T (I + V E^{-1} V^T)^{-1} V E^{-1}. \end{aligned}$$

By direct calculation, $V E^{-1} V^T = \text{diag} \{ \omega_i, i = 1, \dots, n \}$ with ω_i as defined above. The calculation of $V P^{-1} V^T$ follows directly from (2.12).

Proposition 2.2. $(I - V P^{-1} V^T)^{1/2} = \text{diag} \left\{ \left[\frac{1}{1 + \omega_i} \right]^{1/2}, i = 1, \dots, n \right\}$ where ω_i is as above.

Proof. The result follows immediately from Proposition 2.1.

Proposition 2.3. With ω_i as above

$$VP^{-1} = \text{diag} \left\{ \frac{1}{1 + \omega_i}, i = 1, \dots, n \right\} VE^{-1}$$

and

$$P^{-1}V^T = E^{-1}V^T \text{diag} \left\{ \frac{1}{1 + \omega_i}, i = 1, \dots, n \right\}.$$

Proof. The results follow directly from (2.12) and Propositions 2.1 and 2.2.

Proposition 2.4. The right hand side of (2.11) is given by

$$\begin{aligned} & (I - VP^{-1}V^T)^{-1/2} [-G_1 + VP^{-1}(V^TG_1 + DG_2)] \\ &= -\text{diag} \left\{ \left[\frac{1}{1 + \omega_i} \right]^{1/2}, i = 1, \dots, n \right\} G_1 - VE^{-1}DG_2. \end{aligned}$$

Proof. We have that

$$\begin{aligned} & -G_1 + VP^{-1}(V^TG_1 + DG_2) \\ &= -(I - VP^{-1}V^T)G_1 + VP^{-1}DG_2 \\ &= -\text{diag} \left\{ \frac{1}{1 + \omega_i}, i = 1, \dots, n \right\} G_1 + VP^{-1}DG_2. \end{aligned}$$

Now the result follows immediately from Propositions 2.2 and 2.3.

With the above formulas, we can solve (2.11) for s by first calculating E and ω in $O(nm)$ operations and then forming

$$\bar{J} = \text{diag} \left\{ \left[\frac{1}{1 + \omega_i} \right]^{1/2}, i = 1, \dots, n \right\} J$$

and

$$y = -\text{diag} \left\{ [1 + \omega_i]^{1/2}, i = 1, \dots, n \right\} (G_1 - VE^{-1}DG_2).$$

Now rewrite the least squares problem (2.11) as

$$(2.13) \quad \begin{pmatrix} \bar{J} \\ \alpha^{1/2}S \end{pmatrix} s =_2 \begin{pmatrix} y \\ 0 \end{pmatrix}$$

which requires $O(np + nm)$ operations. The solution of (2.13) then involves a QR decomposition of \bar{J} (accomplished by Householder transformations with column pivoting) and then a sequence of plane rotations to eliminate $\alpha^{1/2}S$. The cost for this phase is dominated by the $O(np^2)$ operations for the QR decomposition of \bar{J} .

Using Proposition 2.3 and (2.12), it is easily verified that

$$t = -E^{-1} \left[V^T \text{diag} \left\{ \frac{1}{1 + \omega_i}, i = 1, \dots, n \right\} (G_1 + Js - VE^{-1}DG_2) + DG_2 \right]$$

which is dominated in cost by the $O(np)$ operations needed to form Js and several $O(nm)$ terms. Thus the leading cost of calculating a step for ODR is the same $O(np^2)$ operations needed to do the factorization of an $n \times p$ matrix as in OLS. The only additional costs are a small number of calculations costing $O(nm)$ or $O(np)$ operations.

It may occur to the reader that an efficient QR factorization of the matrix in (2.7) might yield a procedure with the same order of work. By re-ordering the upper 2×2 blocks, one can, indeed, do the factorization of this part in $O(np^2)$ operations. The subsequent elimination of the αS and αT blocks, however, would require $O((nm + p)^2)$ operations for each α . It is for this reason, as well as others, that Schwetlick and Tiller [SchT85] do only a "partial" trust region strategy, i.e., their trust region only applies to the step in the β variables. In some badly scaled problems, however, (e.g., Example 3 in Section 4) the ability to scale and constrain the step in δ is essential to solve the problem.

The above formulas for s and t are used for each α value in (2.5). Thus in order to complete the specification of the algorithm, we need to provide the procedure for computing the trust region parameter α to satisfy (2.4) and to discuss a few miscellaneous details.

More [Mor77], following Hebdon [Heb73] (see also [DenS83]), suggested a procedure for computing α in (2.5) so that

$$\|z(\alpha)\| \approx \tau$$

when $\|z(0)\| > \tau$. This procedure is based on approximating the function

$$\phi(\alpha) = \|z(\alpha)\| - \tau$$

by a rational function $\theta(\alpha) = \gamma/(\mu - \alpha)$ where the constants γ and μ are chosen by making $\theta(\alpha^c) = \phi(\alpha^c)$ and $\theta'(\alpha^c) = \phi'(\alpha^c)$. This results in the iteration

$$\alpha^+ = \alpha^c - \frac{\phi(\alpha^c)}{\phi'(\alpha^c)} \times \frac{\phi(\alpha^c) - \tau}{\tau}$$

which is a modified Newton step for the equation $\phi(\alpha) = 0$. In our case, the derivative of $\phi(\alpha)$ is not as simple to compute as in OLS and thus we opt to compute γ and μ by making $\theta(\alpha^c) = \phi(\alpha^c)$ and $\theta(\alpha^-) = \phi(\alpha^-)$. (α^- is the previous estimate.) Then

$$\alpha^+ = \alpha^c - \frac{\phi(\alpha^c)(\alpha^c - \alpha^-)}{\phi(\alpha^c) - \phi(\alpha^-)} \times \frac{\phi(\alpha^c) + \tau}{\tau}$$

which is clearly a modified secant step for the equation $\phi(\alpha) = 0$. More found it necessary to safeguard his procedure by computing and updating upper and lower bounds on α . Similarly, we maintain such bounds, but with formulas appropriate to the secant-like method. These will be provided in a subsequent paper.

The trust region bound τ is updated according well-tested ideas which are in several existing codes. (See e.g., [Mor77], [Gay84], [SchKW86].) A valuable feature in our code has been the "internal doubling" step. For a given η^c and τ , suppose z_τ is generated such that τ restricts z and the reduction in $\|G\|$ predicted by the linear approximation agrees with the actual reduction in $\|G\|$ to a high precision. Normally, one would accept z_τ , set $\eta^+ := \eta^c + z_\tau$ and double τ for the next step. The internal doubling procedure is to remember $\eta^r := \eta^c + z_\tau$, double τ and generate a new z_τ from η^c . Note that this procedure only costs an evaluation of G and, if successful, may save several evaluations of J . In practice, it has been successful often enough to warrant leaving it in. Its main advantage is that it permits rapid and cheap increases in τ based on an overly conservative initial value of τ or when the iterates are moving away from a highly nonlinear region in parameter space.

Since many users will want to compare the results of OLS with ODR, an option to do OLS has been implemented. Enabling this option merely initializes the δ vector to zero and sets V to zero whenever it is computed. It is easily verified that, in this case, (2.11) reduces to the OLS Levenberg-Marquardt step and (2.9) yields $t = 0$ leaving $\delta = 0$. Using this procedure to do OLS, therefore, is equivalent to a standard OLS algorithm with a moderate extra algebraic overhead.

3. Local and Global Convergence Analysis

The global convergence properties of trust-region-Levenberg-Marquardt methods applied to the general nonlinear least squares problem (2.3) are well known (see e.g., [Pow75], [Mor77], [MorS81], [ShuSB85]). As long as the sequence of Jacobian matrices, $\{G'(\eta_k)\}$, is uniformly bounded, then

$$\lim_{k \rightarrow \infty} G'(\eta_k)^T G(\eta_k) = 0,$$

so that any cluster point satisfies the first order necessary conditions for a local minimizer. These results apply to our algorithm and nothing more needs to be said regarding global convergence.

The local convergence behavior of general trust-region-Levenberg-Marquardt methods for nonlinear least squares is discussed by Byrd and Schnabel [ByrS85] who show that, if there is a cluster point η_* where $G'(\eta_*)$ is nonsingular, then the iterates converge at least linearly to η_* independent of the size of $G(\eta_*)$. This theory also applies to our algorithms. If, in addition, the residual $G(\eta_*)$ is sufficiently small, Byrd and Schnabel show that asymptotically the trust region constraint becomes inactive, and that the Levenberg-Marquardt algorithm reduces to the Gauss-Newton iteration

$$\eta_{k+1} = \eta_k - [G'(\eta_k)^T G'(\eta_k)]^{-1} G'(\eta_k)^T G(\eta_k)$$

and is linearly convergent to η_* . The linear convergence analysis of the Gauss-Newton method is well known (see e.g., [OrtR70], [DenS83]). The constant of linear convergence depends upon the smallest singular value of $G'(\eta_*)$, the residual $G(\eta_*)$, and the nonlinearity of $G(\eta)$ near η_* .

The small residual analysis is particularly relevant to ODR because most applications of ODR will have small residuals. This is especially true when ODR is used to consider

errors in independent variables in parameter estimation, because errors in the independent variables are most likely to be considered when the model and the dependent variable measurements are accurate, which implies that the residuals will be small.

It turns out that the application of the local Gauss-Newton analysis to ODR is non-trivial, although the expected results can be proven. This is the main contribution of this section. To simplify the algebra here, we consider a version of the ODR problem (2.1) with the simplified weighting scheme $w_i = 1$ and $d_i = \sigma$ for all i , i.e.,

$$(3.1) \quad \min_{\beta, \delta} \sum_{i=1}^n \left[(f(x_i + \delta_i; \beta) - y_i)^2 + \sigma^2 \delta_i^T \delta_i \right]$$

where $\sigma > 0$. This weighting still allows the metric of distance from the curve $f(x; \beta)$ to the data points (x_i, y_i) to vary from vertical (as $\sigma \rightarrow \infty$) to orthogonal ($\sigma = 1$) to horizontal (as $\sigma \rightarrow 0$). (We explain this statement more carefully later in this section.) This is all the generality in the weighting that is usually used in practice, and precisely what we use in most of our computational results in Section 4.

In fact, as we illustrate in Section 4, in practical ODR applications, the user may wish to solve (3.1) for various values of σ . A second contribution of this section is to produce a bound on the constant of linear convergence of the Gauss-Newton method applied to (3.1) that is independent of the value of σ .

To further simplify notation, we rewrite (3.1) as

$$(3.2a) \quad \min_{\eta} R(\eta)^T R(\eta) + \sigma^2 \delta^T \delta$$

or equivalently,

$$(3.2b) \quad \min_{\eta} G(\eta)^T G(\eta)$$

where $\delta = (\delta_1^T, \delta_2^T, \dots, \delta_n^T)^T$, $\eta = (\beta^T, \delta^T)^T$, $R(\eta)_i = f(x_i + \delta_i; \beta) - y_i$, $i = 1, \dots, n$, and $G(\eta) = (R(\eta)^T, \sigma \delta^T)^T$. Our analysis will not depend upon the special form of $R(\eta)$ in any way. Recall that

$$G'(\eta) = \begin{pmatrix} J(\eta) & V(\eta) \\ 0 & \sigma I \end{pmatrix}$$

where $J(\eta)$ and $V(\eta)$ are, as in Section 2, the derivatives of $R(\eta)$ with respect to β and δ respectively.

The difficulty in applying standard Gauss-Newton analysis to (3.2) is that $G(\eta)$ and $G'(\eta)$ are functions of σ . In Theorem 3.5 we show that the convergence can be analyzed in terms of the properties of $J(\eta)$, $V(\eta)$, $R(\eta)$, and δ , only, i.e., independent of σ except for its role in determining η . Lemmas 3.3 and 3.4 are used in the proof of Theorem 3.5 to bound the linear and quadratic terms in the convergence results, respectively, independent of σ . Lemmas 3.1 and 3.2 are technical results used in the proof of Lemma 3.3. Theorem 3.5 can be applied to the more generally weighted problem (2.1) by making an appropriate change of variables.

Lemma 3.1. Let $A \in R^{p \times p}$, $C \in R^{q \times q}$, $B \in R^{p \times q}$, and let

$$H = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$$

be symmetric and positive definite. Then $\|H\| \leq \|A\| + \|C\|$.

Proof: Define $a = \|A\|$, $c = \|C\|$. For any $v \in R^p$ and $w \in R^q$, consider

$$y = \begin{bmatrix} v\|w\|\sqrt{c} \\ -w\|v\|\sqrt{a} \end{bmatrix}.$$

Then, since H is positive definite,

$$0 \leq y^T H y = v^T A v \|w\|^2 c - 2v^T B w \|w\| \|v\| \sqrt{ac} + w^T C w \|v\|^2 a$$

and therefore

$$\begin{aligned} 2v^T B w &\leq \frac{v^T A v \|w\|^2 c + w^T C w \|v\|^2 a}{\|w\| \|v\| \sqrt{ac}} \\ &\leq \frac{\|v\|^2 a \|w\|^2 c + \|w\|^2 c \|v\|^2 a}{\|w\| \|v\| \sqrt{ac}} \\ &= 2\|v\| \|w\| \sqrt{ac} \\ &\leq \|v\|^2 c + \|w\|^2 a. \end{aligned}$$

Thus

$$\begin{aligned} \|H\| &= \max_{v \in R^p, w \in R^q} \frac{v^T A v + 2v^T B w + w^T C w}{v^T v + w^T w} \\ &\leq \max_{v \in R^p, w \in R^q} \frac{\|v\|^2 a + (\|v\|^2 c + \|w\|^2 a) + \|w\|^2 c}{v^T v + w^T w} \\ &= a + c. \end{aligned}$$

Lemma 3.2. Let $A \in R^{n \times p}$ have full column rank and $B \in R^{n \times n}$ be positive definite. Then $\|(A^T B^{-1} A)^{-1}\| \leq \|(A^T A)^{-1}\| \|B\|$.

Proof:

$$\begin{aligned} \|(A^T B^{-1} A)^{-1}\| &= \max_{v \in R^p} \frac{v^T v}{v^T A^T B^{-1} A v} \\ &\leq \max_v \frac{v^T v \|B\|}{v^T A^T A v} \leq \|(A^T A)^{-1}\| \|B\|. \end{aligned}$$

Lemma 3.3. Let $J \in R^{n \times p}$ have full column rank, $V \in R^{n \times q}$, I the $q \times q$ identity, and σ a positive scalar. Let $M(\sigma) = \begin{pmatrix} J & V \\ 0 & \sigma I \end{pmatrix}$, and $N(\sigma) = M(\sigma)^T M(\sigma)$. Then $N(\sigma)$ is nonsingular and

$$(3.3) \quad \|N(\sigma)^{-1}\| \leq \|(J^T J)^{-1}\| + \sigma^{-2} (1 + \|(J^T J)^{-1}\| \|V\|^2).$$

Proof: Since J has full column rank, $M(\sigma)$ has full column rank. Thus $N(\sigma)$ is positive definite, and it is straight forward to verify that $N(\sigma)^{-1} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$, where

$$\begin{aligned} A &= [\sigma^2 J^T (V^T V + \sigma^2 I)^{-1} J]^{-1}, \\ C &= [\sigma^2 I + V^T (I - J(J^T J)^{-1} J^T) V]^{-1}, \\ B &= -(J^T J)^{-1} J^T V C. \end{aligned}$$

From Lemma 3.2,

$$\begin{aligned} \|A\| &\leq -\sigma^{-2} \|(J^T J)^{-1}\| \|V^T V + \sigma^2 I\| \\ (3.4) \quad &= \sigma^{-2} \|(J^T J)^{-1}\| (\|V^T V\| + \sigma^2) \\ &= \sigma^{-2} \|(J^T J)^{-1}\| \|V\|^2 + \|(J^T J)^{-1}\|. \end{aligned}$$

Since $C^{-1} - \sigma^2 I = V^T [I - J(J^T J)^{-1} J^T] V$ is positive semi-definite, the smallest eigenvalue of C^{-1} is at least σ^2 which shows

$$(3.5) \quad \|C\| \leq \sigma^{-2}.$$

Thus, since $N(\sigma)^{-1}$ is positive definite, applying Lemma 3.1 and using (3.4) and (3.5) gives (3.3).

Lemma 3.4. Let the assumptions of Lemma 3.3 hold, and let $u \in R^n$. Let z be the solution to

$$(3.6) \quad \min_{z \in R^{p+q}} \left\| \begin{pmatrix} J & V \\ 0 & \sigma I \end{pmatrix} z - \begin{pmatrix} u \\ 0 \end{pmatrix} \right\|.$$

Let $J^+ = (J^T J)^{-1} J^T$ and let V^+ denote the pseudoinverse of V . Then

$$\|z\| \leq [\|V^+\| + \|J^+\| (1 + \|V\| \|V^+\|)] \|u\|.$$

Proof: Define $y \in R^{p+q}$ by

$$z = \begin{pmatrix} I & -J^+ V \\ 0 & I \end{pmatrix} y$$

and let $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$, $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ where $z_1, y_1 \in R^p$, and $z_2, y_2 \in R^q$. Then (3.6) is equivalent to

$$(3.7) \quad \min_{y \in R^{p+q}} \left\| \begin{pmatrix} J & PV \\ 0 & \sigma I \end{pmatrix} y - \begin{pmatrix} u \\ 0 \end{pmatrix} \right\|$$

where $P = (I - JJ^+)$ is a projection matrix. From the normal equations using $J^T p = 0$, the solution to (3.7) is

$$\begin{aligned} y_1 &= J^+ u \\ y_2 &= (V^T P V + \sigma^2 I)^{-1} V^T P u. \end{aligned}$$

It is well known, for instance from the theory of Levenberg-Marquardt methods, that

$$\|y_2\| \leq \|(PV)^+ u\|,$$

and since $(PV)^+ = V^+P$ and $\|P\| \leq 1$,

$$\|y_2\| \leq \|V^+\| \|u\|, \quad \|y_1\| \leq \|J^+\| \|u\|.$$

Finally since $z_2 = y_2$ and $z_1 = y_1 - J^+V y_2$,

$$\|z_2\| \leq \|V^+\| \|u\|$$

and

$$\begin{aligned} \|z_1\| &\leq \|y_1\| + \|J^+\| \|V\| \|y_2\| \\ &\leq \|J^+\| (1 + \|V\| \|V^+\|) \|u\|. \end{aligned}$$

Using $\|z\| \leq \|z_1\| + \|z_2\|$ gives the desired result.

Throughout the statement and proof of Theorem 3.5, we will often omit the argument η ; i.e., we will denote $G(\eta_*)$ and $G(\eta_0)$ by G_* and G_0 , respectively, and likewise for other symbols in place of G . Also for J having full column rank, J^+ will denote $(J^T J)^{-1} J^T$, and for V having full row rank, V^+ will denote $V^T (V V^T)^{-1}$. Note that $\|J^+\|^2 = \|(J^T J)^{-1}\|$.

Theorem 3.5. Let $R(\eta) : R^t \rightarrow R^n$ be continuously differentiable in an open convex set $D \subset R^t$. Let $\eta^T = (\beta^T, \delta^T)$, $\beta \in R^p$, $\delta \in R^q$, let σ be a positive scalar, and let $G(\eta) = \begin{pmatrix} R(\eta) \\ \sigma \delta \end{pmatrix}$. Assume there exists $\eta^* \in D$ such that $G'(\eta)^T G(\eta^*) = 0$, and that there exists $\gamma \geq 0$ for which

$$(3.8) \quad \|R'(\eta) - R'(\eta^*)\| \leq \gamma \|\eta - \eta^*\|$$

for all $\eta \in D$. Define

$$\begin{aligned} c_1 &= \gamma \left[\|(J_*^T J_*)^{-1}\| \|R_*\| + \left(1 + \|(J_*^T J_*)^{-1}\| \|V_*\|^2\right) \|V_*^+\| \|\delta_*\| \right] \\ c_2 &= (\gamma/2) [\|V_*^+\| + \|J_*^+\| (1 + \|V_*\| \|V_*^+\|)]. \end{aligned}$$

If $c_1 < 1$, then for any $c \in (1, 1/c_1)$, there exists $\epsilon > 0$ such that for all η_0 for which $\|\eta_0 - \eta_*\| \leq \epsilon$, the sequence generated by the Gauss-Newton method

$$\eta_{k+1} = \eta_k - (G'(\eta_k)^T G'(\eta_k))^{-1} G'(\eta_k)^T G(\eta_k)$$

is well defined, converges to η_* and obeys

$$(3.9) \quad \|\eta_{k+1} - \eta_*\| \leq c (c_1 + c_2 \|\eta_k - \eta_*\|) \|\eta_k - \eta_*\|.$$

Proof: The crucial remaining part of the proof is to note that since the optimality condition $G'_* G_* = 0$ gives $V_*^T R_* + \sigma^2 \delta_* = 0$, we have

$$\begin{aligned} R_* &= -\sigma^2 (V_* V_*^T)^{-1} V_*^T \delta_* \\ &= -\sigma^2 (V_*^+)^T \delta_*. \end{aligned}$$

so that

$$\|R_*\| \leq \sigma^2 \|V_*^+\| \|\delta_*\|.$$

Thus from (3.3) of Lemma 3.2,

$$\begin{aligned} (3.10) \quad & \left\| ((G'_*)^T G'_*)^{-1} \right\| \|R_*\| \\ & \leq \left\| (J_*^T J_*)^{-1} \right\| \|R_*\| + \sigma^{-2} \left(1 + \left\| (J_*^T J_*)^{-1} \right\| \|V_*\|^2 \right) \|R_*\| \\ & \leq \left\| (J_*^T J_*)^{-1} \right\| \|R_*\| + \left(1 + \left\| (J_*^T J_*)^{-1} \right\| \|V_*\|^2 \right) \|V_*^+\| \|\delta_*\| \\ & = c_1 / \gamma. \end{aligned}$$

The remainder of the proof is by induction. Let c be a fixed constant in $(1, 1/c_1)$. Then there exists $\epsilon_1 > 0$ such that for all η for which $\|\eta - \eta_*\| \leq \epsilon_1$, $J(\eta)$ and $V(\eta)$ have full column and row rank, respectively, and

$$(3.11) \quad \left\| (G'(\eta)^T G'(\eta))^{-1} \right\| \leq c \left\| ((G'_*)^T G'_*)^{-1} \right\|,$$

and

$$\begin{aligned} (3.12) \quad & \|V(\eta)^+\| + \|J(\eta)^+\| (1 + \|V(\eta)\| \|V(\eta)^+\|) \\ & \leq c [\|V_*^+\| + \|J_*^+\| (1 + \|V_*\| \|V_*^+\|)]. \end{aligned}$$

Let $\epsilon = \min \{ \epsilon_1, (1 - cc_1) / (2cc_2) \}$. Then at the first iteration $((G'_0)^T G'_0)$ is nonsingular, and

$$\begin{aligned} (3.13) \quad \eta_1 - \eta_* &= \eta_0 - \eta_* - ((G'_0)^T G'_0)^{-1} (G'_0)^T G_0 \\ &= a + b \end{aligned}$$

where

$$\begin{aligned} a &= -((G'_0)^T G'_0)^{-1} (G'_0)^T G_* \\ b &= ((G'_0)^T G'_0)^{-1} (G'_0)^T (G_* - G_0 - G'_0(\eta_* - \eta_0)). \end{aligned}$$

Since $(G'_*)^T G_* = 0$,

$$(3.14) \quad (G'_0)^T G_* = (G'_0 - G'_*)^T G_* = (R'_0 - R'_*)^T R_*.$$

due to the linearity of the final q components of $G(\eta)$. Using (3.14) plus (3.8), (3.10), and (3.11), we have

$$\begin{aligned}
 (3.15) \quad \|a\| &\leq \left\| ((G'_0)^T G'_0)^{-1} \right\| \|R'_0 - R'_*\| \|R_*\| \\
 &\leq c\gamma \left\| ((G'_*)^T G'_*)^{-1} \right\| \|R_*\| \|\eta_0 - \eta_*\| \\
 &\leq cc_1 \|\eta_0 - \eta_*\|.
 \end{aligned}$$

Also note that b is the solution to

$$\min \|G'_0 b - [G_* - G_0 - G'_0(\eta_* - \eta_0)]\|,$$

and that

$$(3.16) \quad G_* - G_0 - G'_0(\eta_* - \eta_0) = \begin{pmatrix} R_* - R_0 - R'_0(\eta_* - \eta_0) \\ 0 \end{pmatrix}.$$

From (3.8) and standard results

$$(3.17) \quad \|R_* - R_0 - R'_0(\eta_* - \eta_0)\| \leq (\gamma/2) \|\eta_0 - \eta_*\|^2.$$

Thus using (3.16), (3.17), Lemma 3.4, and (3.12),

$$\begin{aligned}
 (3.18) \quad \|b\| &\leq [\|V_0^+\| + \|J_0^+\| (1 + \|V_0\| \|V_0^+\|)] (\gamma/2) \|\eta_0 - \eta_*\|^2 \\
 &\leq cc_2 \|\eta_0 - \eta_*\|^2.
 \end{aligned}$$

Substituting (3.15) and (3.18) into (3.13) and recalling $\|\eta_0 - \eta_*\| \leq (1 - cc_1)/2cc_2$,

$$\begin{aligned}
 \|\eta_1 - \eta_*\| &\leq c(c_1 + c_2 \|\eta_0 - \eta_*\|) \|\eta_0 - \eta_*\| \\
 &\leq [(1 + cc_1)/2] \|\eta_0 - \eta_*\|
 \end{aligned}$$

which proves (3.9) in the case $k = 0$ and shows that $\|\eta_1 - \eta_*\| < \|\eta_0 - \eta_*\|$. The proof of the induction step is identical.

Byrd and Schnabel [ByrS85] show that the trust-region-Levenberg-Marquardt algorithm described in Section 2 will reduce asymptotically to Gauss-Newton if the constant of linear convergence, c_1 , in Theorem 3.5 is sufficiently small. In particular, our computer code will accept the trial step if the ratio of actual to "predicted" reduction,

$$\frac{\|G(\eta_c + z_r)\| - \|G(\eta_c)\|}{\|G(\eta_c) + G'(\eta_c)z_r\| - \|G(\eta_c)\|}$$

is at least 0.001. In this case, as long as $c_1 < .9999$, the trust region in the Levenberg-Marquardt algorithm is inactive asymptotically, so that the algorithm becomes Gauss-Newton.

Now we consider the behavior of the ODR problem (3.2) as the parameter σ is varied. For this purpose, let us denote the global minimizer to (3.2) by $\eta_*(\sigma)$. Then by standard analyses of barrier function methods, (see e.g., [FiaM68] or [Lue73]) we know that the limit of $\eta_*(\sigma)$ as $\sigma \rightarrow \infty$ is the solution to

$$\min_{\eta} \|R(\eta)\|^2 \quad \text{subject to } \delta = 0,$$

i.e., the standard OLS problem

$$(3.19) \quad \min_{\beta} \|R(\beta, 0)\|^2.$$

Similarly, the limit of $\eta_*(\sigma)$ as $\sigma \rightarrow 0$ is the solution to the implicit least squares (ILS) problem

$$(3.20) \quad \min_{\eta} \|\delta\|^2 \quad \text{subject to } R(\eta) = 0.$$

In the data fitting context where $R(\eta)_i = f(x_i + \delta_i; \beta) - y_i$, (3.19) is the standard problem where the independent variables x_i are assumed exact so that the metric of distance is in the y (vertical) direction only. In contrast (3.20) is the case where the dependent variables y_i are assumed exact and the independent variables x_i inexact, so that the metric is entirely in the x (horizontal) direction.

The standard analysis of barrier function methods also shows that $\|R(\eta_*(\sigma))\|$ is a monotonically increasing function of σ , and that $\|\delta_*(\sigma)\|$ is a monotonically decreasing function of σ . This means that for all $\sigma \in (0, \infty)$, the values of $\|R(\eta_*(\sigma))\|$ and $\|\delta_*(\sigma)\|$ are bounded above by the optimal objective function values for problems (3.19) and (3.20), respectively. In data fitting terms, for any σ , the norm of the optimal vertical residuals in ODR is bounded above by the norm of the optimal residuals in OLS, and the norm of the optimal horizontal residuals in ODR is bounded above by the norm of the optimal residual for the ILS problem. The computational results of Section 4 demonstrate these relationships.

Combining the above facts with Theorem 3.5 shows that, if the optimal objective function values for problems (3.19) and (3.20) are sufficiently small, and if $J(\eta_*(\sigma))$ and $V(\eta_*(\sigma))$ are sufficiently well-conditioned for all $\sigma \in (0, \infty)$, then the Gauss-Newton algorithm applied to (3.2) is linearly convergent for any $\sigma \in (0, \infty)$.

Corollary 3.6. Let η , β , δ , $R(\eta)$, $G(\eta)$, $J(\eta)$, and $V(\eta)$ be defined as in Theorem 3.5. For any $\sigma \in (0, \infty)$, let $\eta_*(\sigma) = (\beta_*(\sigma)^T, \delta_*(\sigma)^T)^T$ denote the global solution to

$$(3.21) \quad \min_{\beta, \delta} \|R(\beta, \delta)\|^2 + \sigma^2 \|\delta\|^2.$$

Also let β_{OLS} denote the global solution to the ordinary least squares problem

$$\min_{\beta} \|R(\beta, 0)\|^2$$

and let $(\beta_{ILS}, \delta_{ILS})$ denote the global solutions to the implicit least squares problem

$$\min_{\beta, \delta} \|\delta\|^2 \text{ subject to } R(\beta, \delta) = 0.$$

Let $R_{OLS} = R(\beta_{OLS})$. Assume that there exist $\hat{\epsilon} > 0$, $\hat{\gamma} \geq 0$ such that for each $\sigma \in (0, \infty)$,

$$\|R'(\eta) - R'(\eta_*(\sigma))\| \leq \hat{\gamma} \|\eta - \eta_*(\sigma)\|$$

for all η for which $\|\eta - \eta_*(\sigma)\| \leq \hat{\epsilon}$. Assume also that for all $\sigma \in (0, \infty)$, $J(\eta_*(\sigma))$ and $V(\eta_*(\sigma))$ have full column and row rank, respectively, and let \hat{J} , \hat{J}^+ , \hat{V} , and \hat{V}^+ be uniform bounds on $\|J(\eta_*(\sigma))\|$, $\|J(\eta_*(\sigma))^+\|$, $\|V(\eta_*(\sigma))\|$, and $\|V(\eta_*(\sigma))^+\|$, respectively, over all $\sigma \in (0, \infty)$. Define

$$\begin{aligned} \hat{c}_1 &= \hat{\gamma} \left[\left(\hat{J}^+ \right)^2 R_{OLS} + \left(1 + \left(\hat{J}^+ \right)^2 \hat{V}^2 \right) \hat{V} \delta_{ILS} \right] \\ \hat{c}_2 &= (\hat{\gamma}/2) \left[\hat{V}^+ + \hat{J}^+ \left(1 + \hat{V} \hat{V}^+ \right) \right]. \end{aligned}$$

If $\hat{c}_1 < 1$, then for any $c \in (1, 1/\hat{c}_1)$, there exists $\hat{\epsilon} > 0$ such that for any $\sigma \in (0, \infty)$, the sequence $\{\eta_k\}$ generated by the Gauss-Newton method applied to (3.2) starting from any η_0 for which $\|\eta_0 - \eta_*(\sigma)\| \leq \hat{\epsilon}$ is well-defined, converges to $\eta_*(\sigma)$, and obeys

$$\|\eta_{k+1} - \eta_*(\sigma)\| \leq c [\hat{c}_1 + \hat{c}_2 \|\eta_k - \eta_*(\sigma)\|] \|\eta_k - \eta_*(\sigma)\|.$$

4. Computational Testing

In this section we report the results of preliminary computational testing. These tests, consisting of two contrived problems and one real problem, were selected in order to illustrate the effectiveness of the implementation and to demonstrate the performance of the basic algorithm. They also allow us to contrast ODR and OLS, which can have rather dramatic differences, and to point out some of the inherent difficulties in ODR problems.

The contrast between OLS and ODR is best brought out in terms of the parameter σ and the function $\beta(\sigma)$ from Section 3. (Recall that $\beta(\infty)$ corresponds to the OLS solution.) Since, in practice, the correct value of σ may not be known exactly, it is of interest to compute $\beta(\sigma)$ for various values of σ .

The algorithm was coded in Fortran 77 and run in double precision on the Perkin-Elmer 3230 at the National Bureau of Standards (NBS). The graphics were done on the Evans and Sutherland PS-300, also located at NBS.

Example 1.

Consider

$$y = \frac{1}{x-1}$$

and define $x_i = .01 + (i-1) * .05$, $i = 1, \dots, 40$. Next let

$$y_i = \frac{1}{x_i - 1}, \quad i = 1, \dots, 40.$$

Now we perturb the data points as follows:

$$x_i := x_i + rx$$

$$y_i := y_i + ry$$

where the rx are uniformly distributed on $(-.05, .05)$ and the ry are uniformly distributed on $(-.25, .25)$. The model for the data was taken to be

$$y = \frac{\beta_1}{x - \beta_2}$$

and the ODR program was run with several values of σ . The results are reported in two tables and three graphs. Table 1 was generated by setting $\sigma = 1$ and taking $\beta^0 = (1, 1)^T$. Subsequent solutions for higher values of σ used the previous solution for the initial approximation. In addition to the values of $\beta(\sigma)$, Table 1 contains the number of evaluations of the extended residual function G (cf (2.3)) and its Jacobian, and the optimal values of $\|R(\eta(\sigma))\|$ and $\|\delta(\sigma)\|$ for each value of σ . Since the value of δ was expected to be approximately the size of the variance of the errors, we set the weight $T = 10$. Table 2 is organized just as Table 1, but the results were generated by starting with the OLS solution using $\beta^0 = (1, 1)^T$ and then decreasing σ . The graphs are as follows: Figure 5 corresponds to the $\sigma = 2$ fit; Figure 6 corresponds to the OLS fit from Table 1; and Figure 7 to the OLS fit from Table 2. Note that on all of these plots, the y-axis has been scaled by a factor of approximately 100 in order to get all of the points on the plot. Because of this scaling, the error pegs (connecting the data points to their predicted values) appear to be much more horizontal than they really are.

Obviously, Tables 1 and 2 exhibit a nonuniqueness of the solutions. It appears that there are two local solutions for the OLS problem corresponding to the asymptote to $+\infty$ being on the left or right half of the curve, and that the trajectories emanating from these solutions come together around $\sigma = 600$ or that the trajectory represented in Table 2 fails to be continuous near $\sigma = 600$. A possible means of investigating this phenomenon is to write the differential equation describing the trajectory $\beta(\sigma)$ and to study possible bifurcation points. This is not pursued here.

Observe that β_2 determines the location of the asymptote and thus the data locate this parameter very well. Note, however, that the data point near the asymptote, corresponding to $(1.01, 100)^T$, completely dominates the fitting process for OLS in Table 2 and results in a value of $-.3180$ for β_1 . The ODR fit is not nearly so influenced by this data point and, for a broad range of σ , does a very good job of fitting the data. This last point is important, namely that the parameter values do not vary much as a function of σ , which means that σ may not need to be known with much accuracy. The stability of $\beta(\sigma)$ has been noticed on all of our examples and on problems not reported here. This is not, of course, a proof that this phenomenon holds more generally.

A further difference between the OLS and the ODR fits is that the errors for both the the OLS fits do not appear to be random. Almost all of the errors to the left of the asymptote are negative while all to the right are positive. The ODR errors for reasonable values of σ appear to be much more random.

An examination of the computational results reveals that the only hard optimization problem in each set is the first. Subsequent solutions are found very quickly except, of course, for the problem corresponding to $\sigma = 500$ in Table 2 which appears to have jumped across a discontinuity in $\beta(\sigma)$. A detailed examination of the iteration process shows that the algorithm sometimes slows down (a very small value of τ is generated) but then recovers and final convergence is with full Gauss-Newton steps. For the case $\sigma = 500$, fairly large steps in δ were generated which led to apparent convergence with very poor β values (near $(0,1)$) and very large values of δ ($O(1)$). In this case, a very small value of τ was produced. When the procedure was restarted with a large value of τ , the algorithm immediately stepped over this bad region and converged quickly to the correct solution. Thus, it appears to be important to scale the step in δ correctly and to be on the lookout for unrealistic solutions.

Example 2.

This example is a two dimensional version of Example 1. Here we take $x \in R^2$ and

$$y = \frac{1}{x_1 + x_2 - 1}.$$

This function has a line of singularities along $x_1 + x_2 = 1$. We take the data to be on the rectangular grid of width .1 in the x_1 direction and width .2 in the x_2 direction. The first point is $(.01, .01)^T$ and there are 10 points in the x_1 direction and 5 points in the x_2 direction. y is the evaluated at these points and the data are then perturbed according to the following:

$$\begin{aligned}(x_1)_i &:= (x_1)_i + rx \\ (x_2)_i &:= (x_2)_i + rx \\ y_i &:= y_i + ry\end{aligned}$$

where rx are normally distributed with mean 0 and standard deviation .01 and the ry are distributed normally with mean 0 and standard deviation .04.

The form of the model is

$$y = \frac{\beta_1}{\beta_2 x_1 + \beta_3 x_2 - 1}.$$

The results are given in Table 3 and in Figures 8-12. Table 3 is organized as Table 1 and Figures 8-12 correspond to the values of σ noted. The graphs are compressed by a factor of approximately 80 in the y direction in order to get all of the data points on the picture. Again we observe that the values of $\beta(\sigma)$ do not vary quickly and that the fits depend more and more on the points near the asymptote as σ increases. Note that the location of the asymptote is well-determined by the data and that only β_1 changes much as σ increases. Here the insistence on near vertical measures of the error forces β_1 to assume smaller values which has the effect of flattening the function as much as possible near the asymptote. This, of course, tends to minimize the vertical component of the error. As in Example 1, the errors for the OLS fit do not appear random while those for the ODR fits do.

Note that the first solution, corresponding to $\sigma = 1$, was computed with some difficulty. (This is the same situation as occurred for $\sigma = 500$ from Table 2.) In these cases, the terrain in parameter space (β and δ) appears rather flat and fairly large values of δ were again obtained on intermediate iterations. The iteration stalled with an indication of convergence due to x-convergence and a very small value of the trust region radius. A restart (which resets the trust region radius to a larger value) then allows the iterates to step over this flat area and converge very quickly to the correct answer.

The non-uniqueness observed in Example 1 was again observed here. The details are not reported, but we found a second OLS solution which led to a trajectory of solutions that finally joined the above trajectory at $\sigma = 2$.

Example 3.

The data here are actual measurements from a calibration run on an electronic device which was intended to give a flat response over a wide range of frequencies. In the (x, y) -data, the x -values are in units of frequency squared and the y -data are the gain. The x -values are scaled to the interval $(0, 1)$ with several measurements made in each decade from 10^{-8} to 1. More measurements were taken at the higher frequencies since most of the important information is obtained there. The data are plotted in Figures 13 and 14 with a log scale on the x -axis in order to see the situation better. The y -scale on these graphs has been magnified to accentuate the differences. The peak at the right side of the data is at 1.001 while the low point at the extreme right end is at .9473. The flat part at the left side is all near .9882.

The model for this data was obtained from theoretical considerations and has the form

$$y_i = \sum_{j=1}^4 \frac{\alpha_j}{x_i + \gamma_j} + \mu, \quad i = 1, \dots, 44$$

where the parameters to be determined are

$$\beta = (\alpha_1, \dots, \alpha_4, \mu, \gamma_1, \dots, \gamma_4)^T.$$

Estimates of the pole locations—the negative γ -values—are likewise obtained from other analyses. The γ -values are approximately

$$\gamma_1 = 1.38 \times 10^{-3}$$

$$\gamma_2 = 5.96 \times 10^{-2}$$

$$\gamma_3 = 6.71 \times 10^1$$

$$\gamma_4 = 1.07 \times 10^9.$$

Since all of the poles are negative and all of the data have positive x -values, there is no problem with being close to the asymptotes. The range of the x -values, however, implies the need to scale the trust region. We used, for the diagonal scaling matrices S and T , the following:

$$s_i = \frac{1}{|(\beta^0)_i|}$$

$$t_i = \frac{1}{|x_i|}.$$

It turns out that the measurements are proportionately more accurate at the lower frequencies and we therefore took the d -weights to be the same as the t -weights.

While the data were measured quite accurately, there were simply no data at a sufficiently high frequency to warrant keeping the two terms corresponding to $j = 3$ and $j = 4$ in the model. This situation was evidenced by the fact that the Jacobian J had five almost identical columns.

With these terms removed, the resulting problem was easily solved as follows. Using a feature of the program which allows certain parameters to be held fixed at specified values, we fixed the pole values (the γ -values) and used an OLS estimate of the remaining linear parameters. We then freed all of the parameters and did an OLS fit and ODR fits with several values of σ . In doing the ODR fits, we first specified a σ -value of .01 since the gain measurements in this data set were 100 times more accurate than the frequency measurements. Other values of σ were subsequently used for comparison. The results are in Table 4 and Figures 13 and 14 which depict the OLS fit and the ODR fit for $\sigma = .01$, respectively. Virtually no difference appears between the two at the lower frequencies, but some differences occur at the higher frequencies. In the enlargements (Figures 15 and 16) one can easily see that the contribution of the error in the x -values causes ODR to get a significantly better fit than OLS. While the β -values are not reported here, there were, again, very slow changes in $\beta(\sigma)$.

In this section we have shown that our algorithm is effective on highly nonlinear problems, but that these problems themselves often have multiple solutions and other difficulties which imply that potential solutions need to be studied carefully. In subsequent papers, we will provide a more complete description of our implementation and further results on its performance.

Acknowledgements. The authors wish to thank E. Bromberg and T. Griffin of the Center for Applied Mathematics, National Bureau of Standards, for their help in the graphical aspects of these experiments and T. Griffin for his help in the writing of the data handling and driver routines.

Table 1

σ	$\beta_1(\sigma)$	$\beta_2(\sigma)$	Evaluations of		Final Value of	
			G	G'	$\ R(\eta(\sigma))\ $	$\ \delta(\sigma)\ $
1	1.023	1.006	70	25	0.223	0.355
2	1.021	1.005	6	5	0.454	0.223
5	1.015	1.004	6	4	0.771	0.128
25	0.9847	1.002	6	5	1.280	0.080
100	0.9247	0.9972	9	8	3.204	0.061
300	0.9881	0.9928	13	12	10.408	0.035
500	0.9487	0.9953	12	11	15.524	0.018
1000	0.8248	0.9937	10	9	18.881	0.007
∞_1	0.6867	0.9909	7	6	21.774	0.

Table 2

σ	$\beta_1(\sigma)$	$\beta_2(\sigma)$	Evaluations of		Final Value of	
			G	G'	$\ R(\eta(\sigma))\ $	$\ \delta(\sigma)\ $
∞_2	-0.3170	1.010	40	22	104.709	0.
1000	-0.3355	1.095	20	15	104.223	0.007
700	-0.3845	1.093	27	21	103.660	0.015
500*	0.9487	0.9953	103	43	15.524	0.018

Table 3

σ	$\beta_1(\sigma)$	$\beta_2(\sigma)$	$\beta_3(\sigma)$	Evaluations of		Final Value of	
				G	G'	$\ R(\eta(\sigma))\ $	$\ \delta(\sigma)\ $
1*	0.8988	0.9482	1.015	147	60	0.184	0.670
2	0.9223	0.9478	1.019	7	6	0.428	0.618
4	0.9345	0.9506	1.027	8	7	0.989	0.540
10	0.9049	0.9510	1.047	9	8	2.379	0.429
40	0.7148	0.9568	1.044	10	9	6.411	0.315
100	0.3645	0.9343	0.9894	22	16	19.934	0.174
500	0.0914	0.8830	0.9675	25	17	30.424	0.039
∞	0.1192	0.8883	0.9338	27	12	77.440	0.

Table 4

σ	Evaluations of		Final Value of	
	G	G'	$\ R(\eta(\sigma))\ $	$\ \delta(\sigma)\ $
∞	19	9	1.8702	0.
.01	18	7	0.0005	0.03993
.1	12	7	0.0016	0.00482
1.	5	4	0.0018	0.00006

REFERENCES

- [Bar74] Bard, Y. (1974), *Nonlinear Parameter Estimation*, Academic Press, New York.
- [BriL73] Britt, H. I., and R. H. Luecke (1973), "The estimation of parameters in non-linear, implicit models," *Technometrics* 15, 233-247.
- [ByrS85] Byrd, R. H., and R. B. Schnabel (1985), "A unified local and global convergence analysis of Levenberg-Marquardt methods," (in preparation).
- [DenGW81] Dennis, J. E., Jr., D. M. Gay, and R. E. Welsch (1981), "An adaptive nonlinear least-squares algorithm," *TOMS* 7, 348-368.
- [DenS83] Dennis, J. E., Jr., and R. B. Schnabel (1983), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice Hall, Englewood Cliffs, N.J.
- [FiaM68] Fiocco, A. V., and G. P. McCormick (1968), *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York.
- [Ful86] Fuller, W. A. (1986) *Measurement Error Models*, John Wiley, New York.
- [Gay84] Gay, D. M. (1983) "Algorithm 611. Subroutines for unconstrained minimization using a model/trust-region approach," *TOMS* 9, 503-524.
- [GolV83] Golub, G. H., and C. Van Loan (1983), *Matrix Computations*, The Johns-Hopkins University Press.
- [Heb73] Hebden, M. D. (1973), "An algorithm for minimization using exact second derivatives," Rept. TP515m A.E.R.E., Harwell, England.
- [KenSO83] Kendall, M. G., A. Stewart, and J. K. Ord (1983), *The Advanced Theory of Statistics, Fourth Edition*, MacMillan Publishing, New York.
- [LawH74] Lawson, C. L., and R. J. Hanson (1974), *Solving Least Squares Problems*, Prentice Hall, Englewood Cliffs, N.J.
- [Lue73] Luenberger, D. G. (1973), *Introduction to Linear and Nonlinear Programming*, Addison Wesley, Reading, Massachusetts.
- [Mor71] Moran, T. A. P. (1971), "Estimating Structural and functional relationships," *J. Mult. Anal.*, 1, 232-255.
- [Mor77] Moré, J. J. (1977), "The Levenberg-Marquardt algorithm: implementation and theory," in *Numerical Analysis*, G. A. Watson, ed., Lecture Notes in Math. 630, Springer Verlag, Berlin, 105-116.
- [MorS81] Moré, J. J., and D. C. Sorenson (1981), "Computing a trust region step," *SIAM J. Sci. Stat. Comput.* 4, 553-572.
- [OrtR70] Ortega, J. M., and W. C. Rheinboldt (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.
- [Pow75] Powell, M. J. D. (1975), "Convergence properties of a class of minimization algorithms," in *Nonlinear Programming* 2, O. Mangasarian, R. Meyer, and S. Robinson, eds., Academic Press, New York, 1-27.
- [SchKW86] Schnabel, R. B., J. E. Koontz, and B. E. Weiss (1986), "A modular system of algorithms for unconstrained minimization," to appear in *TOMS*.

- [SchSB85] Schultz, G. A., R. B. Schnabel, and R. H. Byrd (1985), "A family of trust-region algorithms for unconstrained minimization with strong global convergence properties," *SIAM J. Numer. Anal.* **22**, 47-67.
- [SchT85] Schwetlick, H., and V. Tiller (1985), "Numerical methods for estimating parameters in nonlinear models with errors in the variables," *Technometrics* **27**, 17-24.
- [Ste73] Stewart, G. W., III (1973) *Introduction to Matrix Computations*, Academic Press, New York.

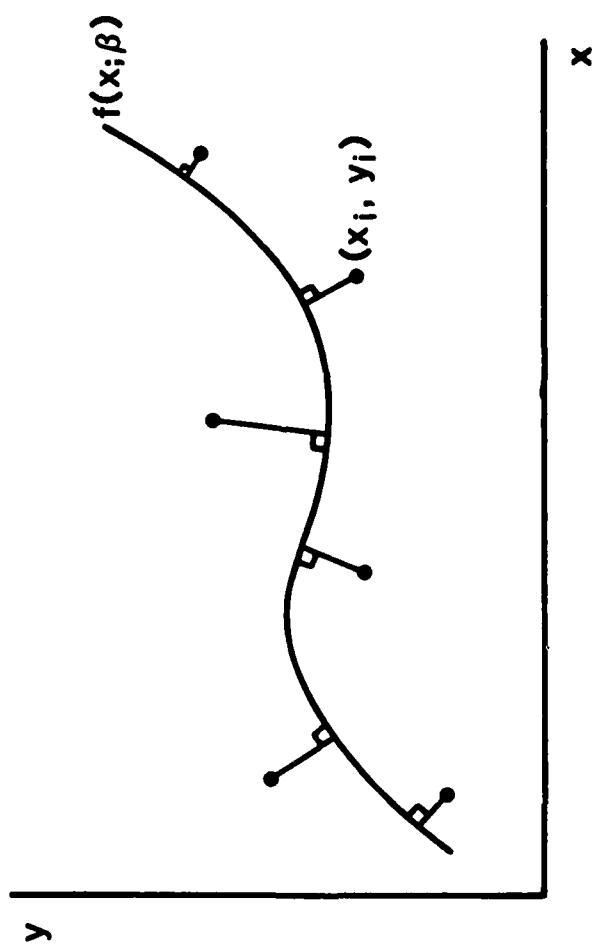


Figure 1

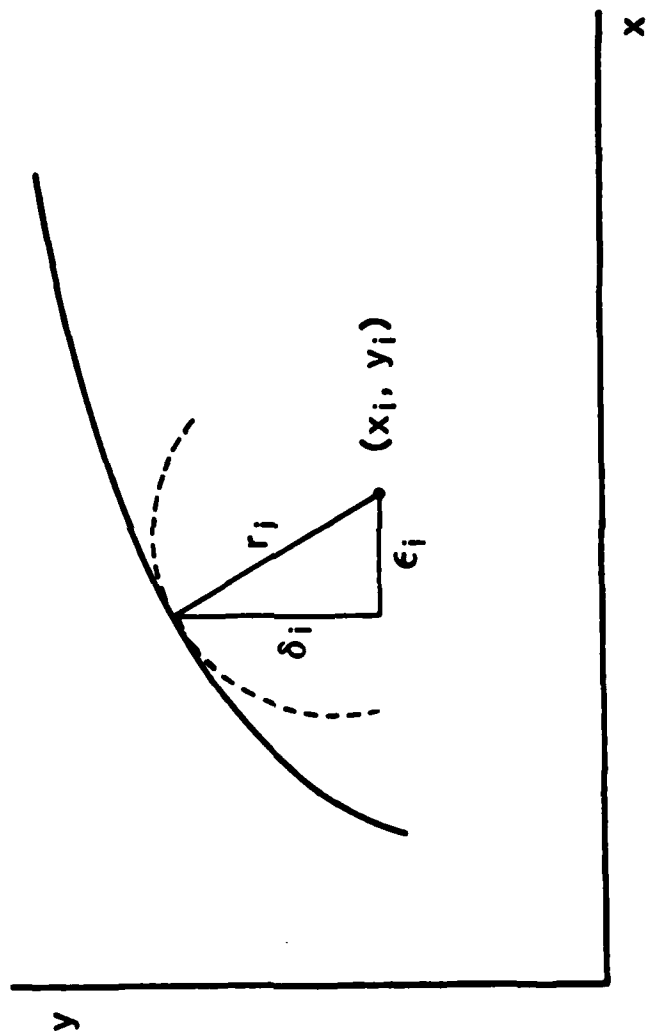


Figure 2

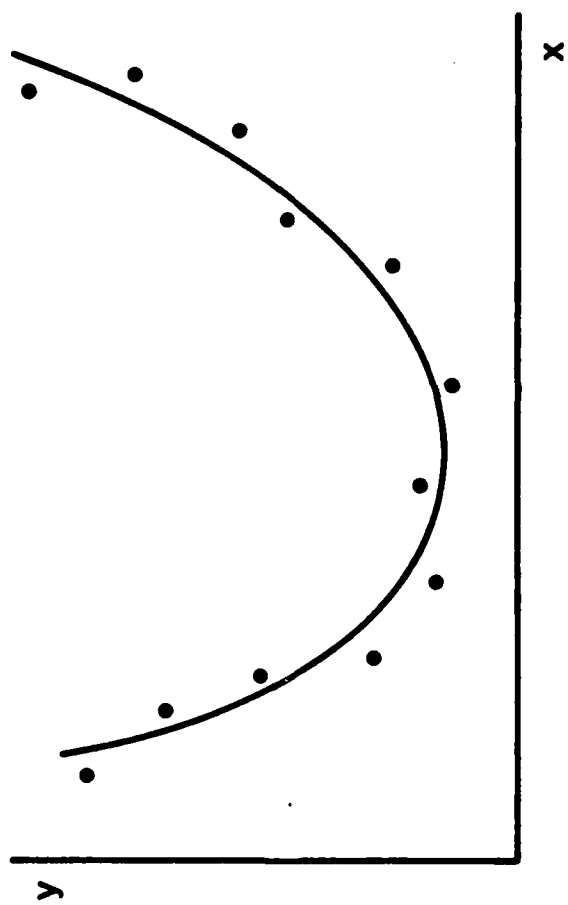


Figure 3

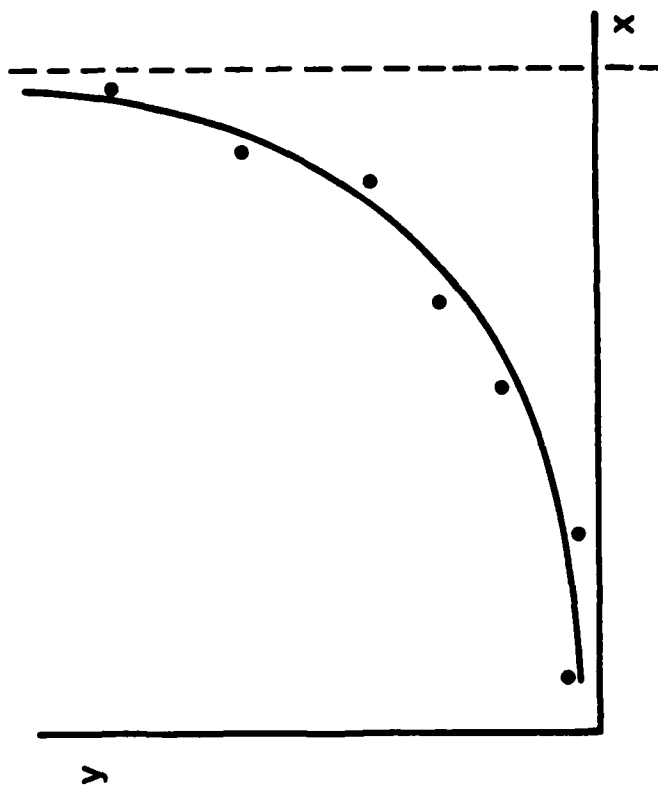


Figure 4

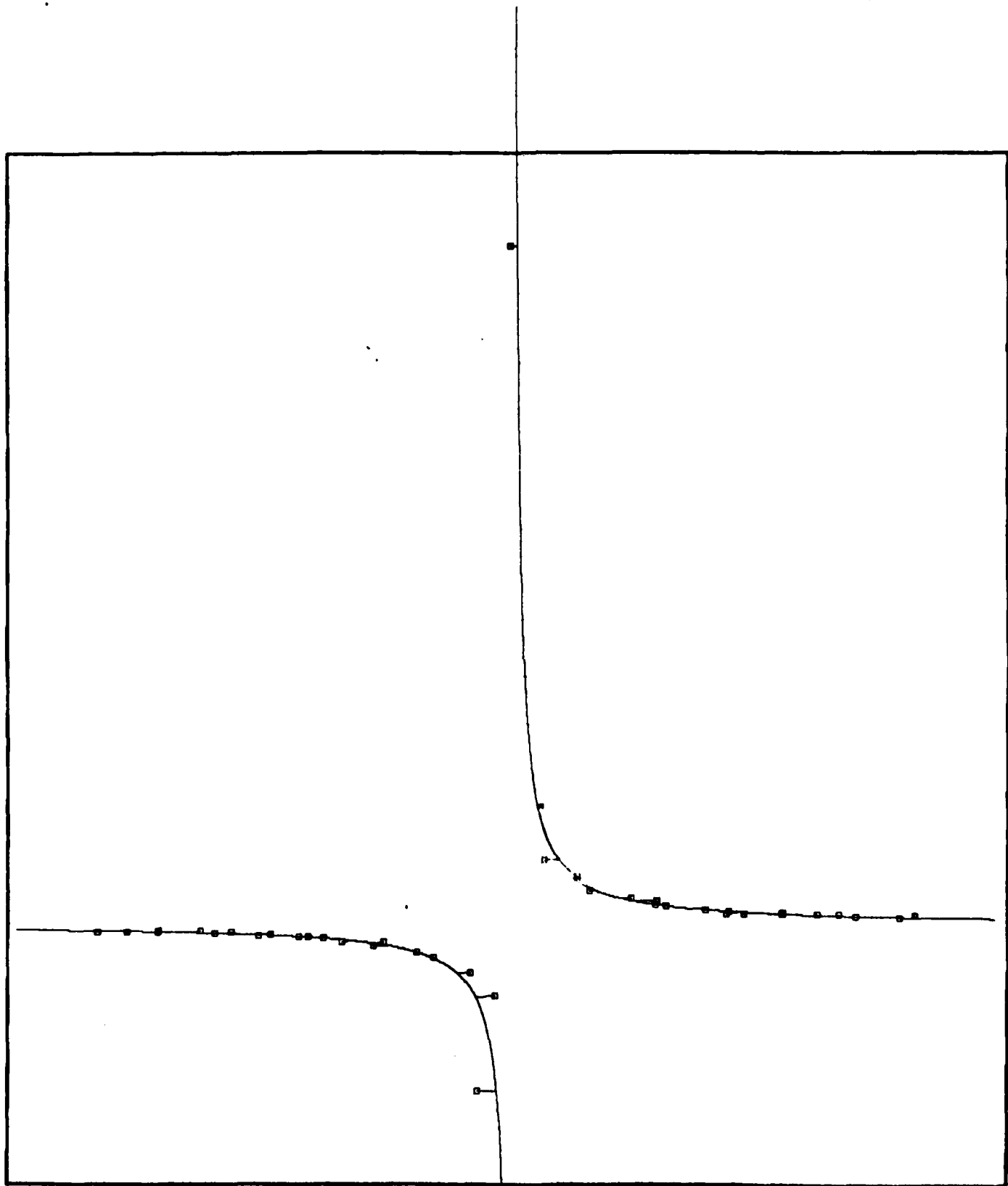


Figure 5

$$a = 5$$

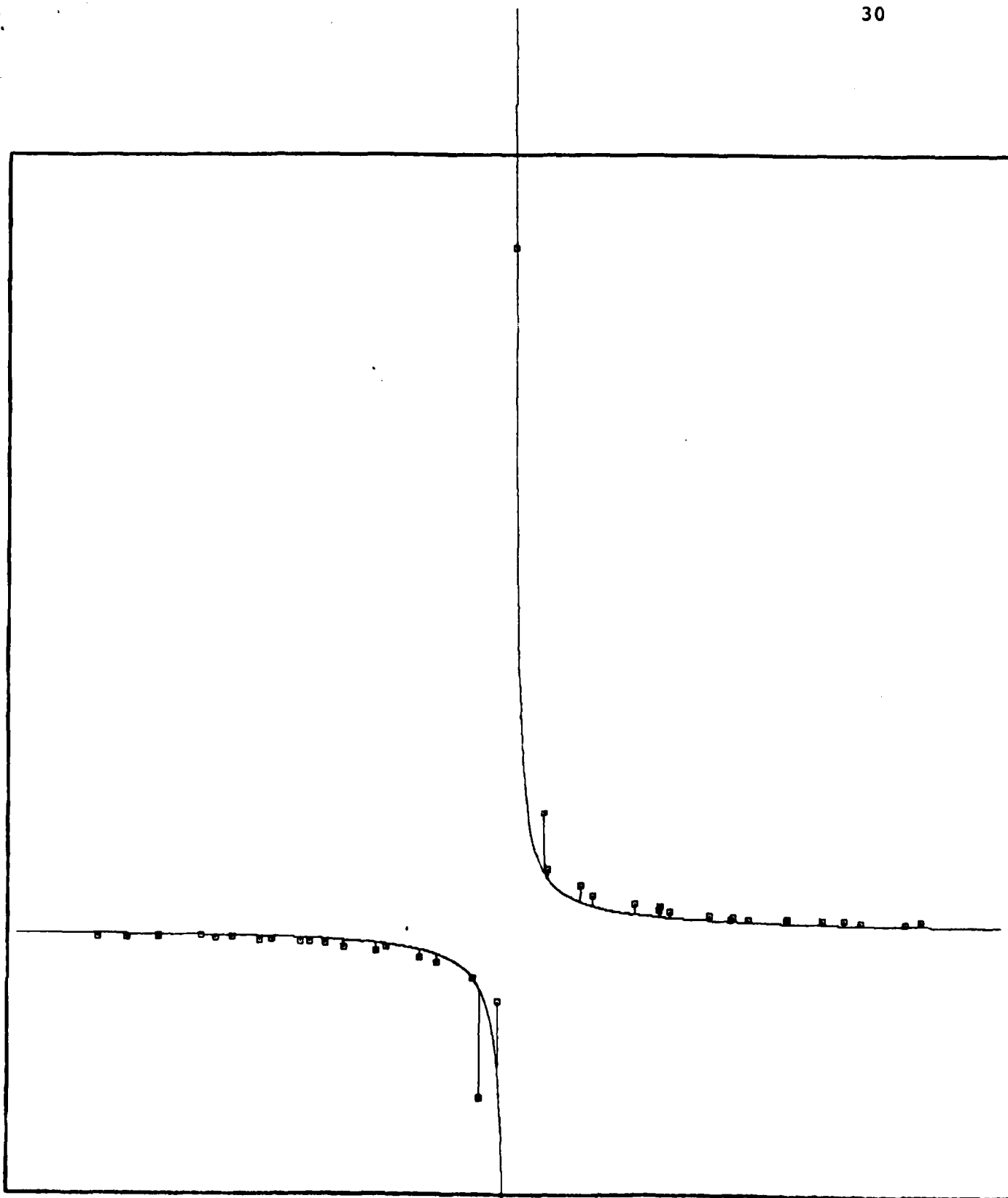


Figure 6
OLS from Table 1

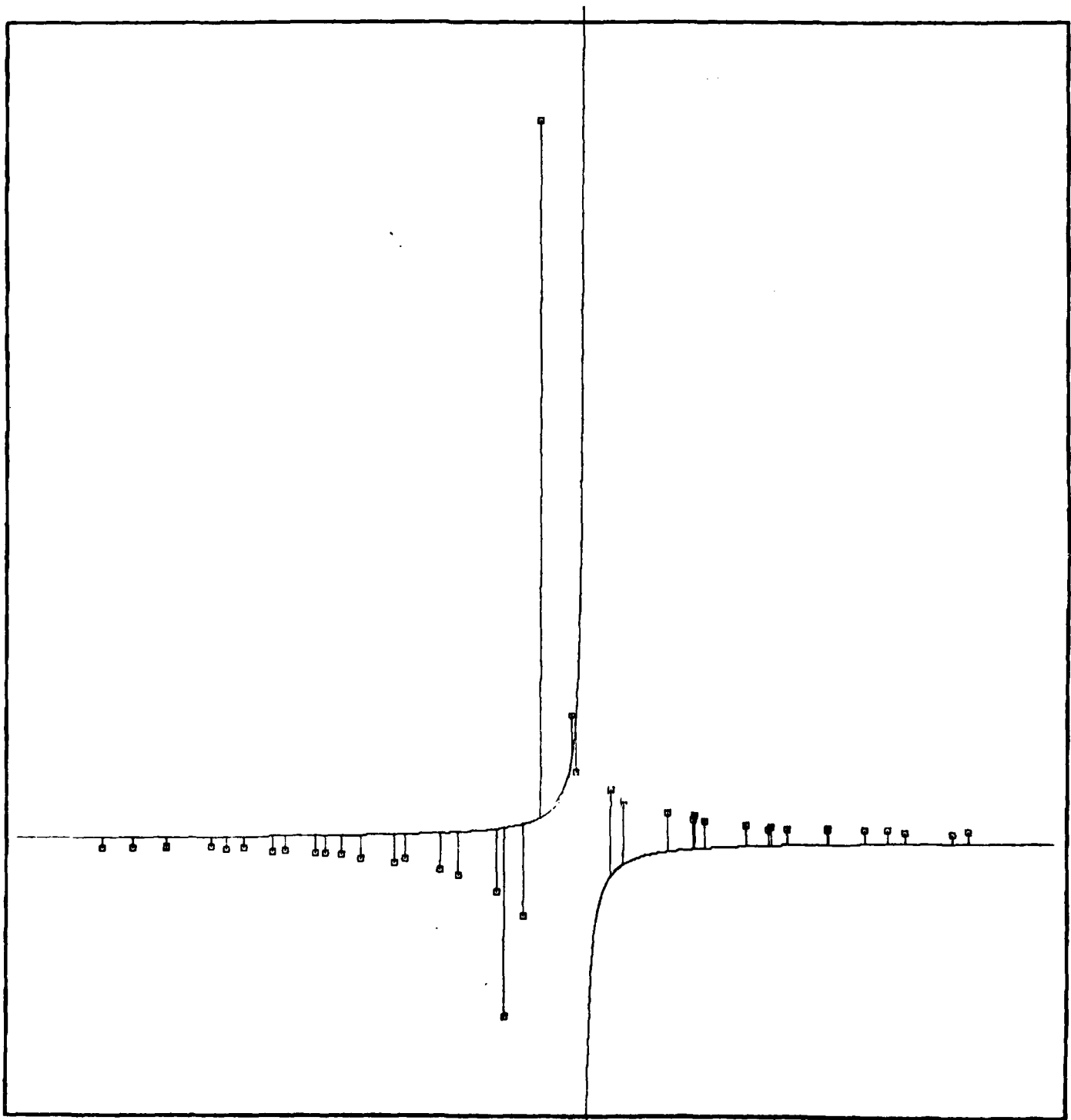


Figure 7
OLS from Table 2

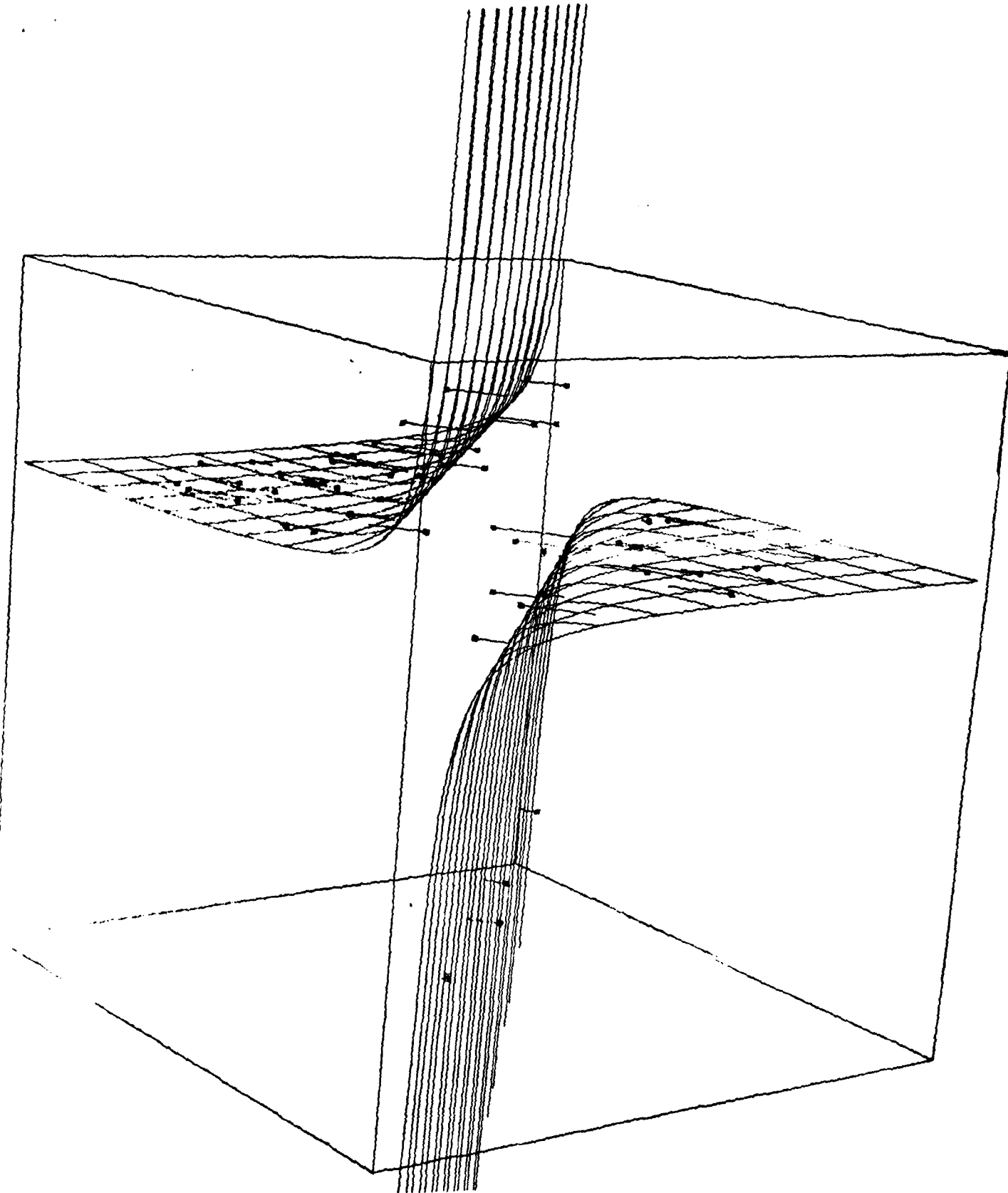


Figure 8

$$\sigma = 2$$

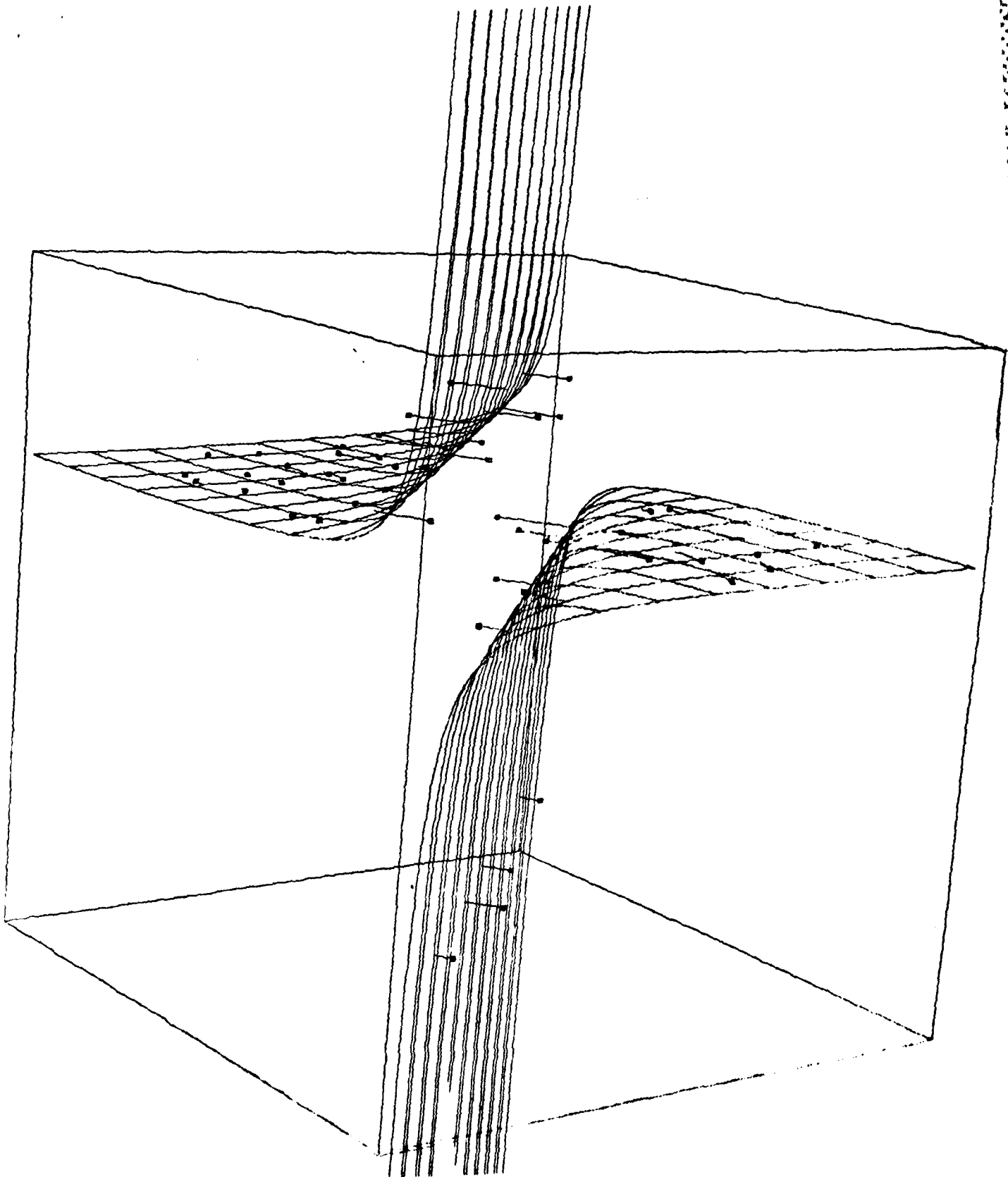


Figure 9
 $\sigma = 10$

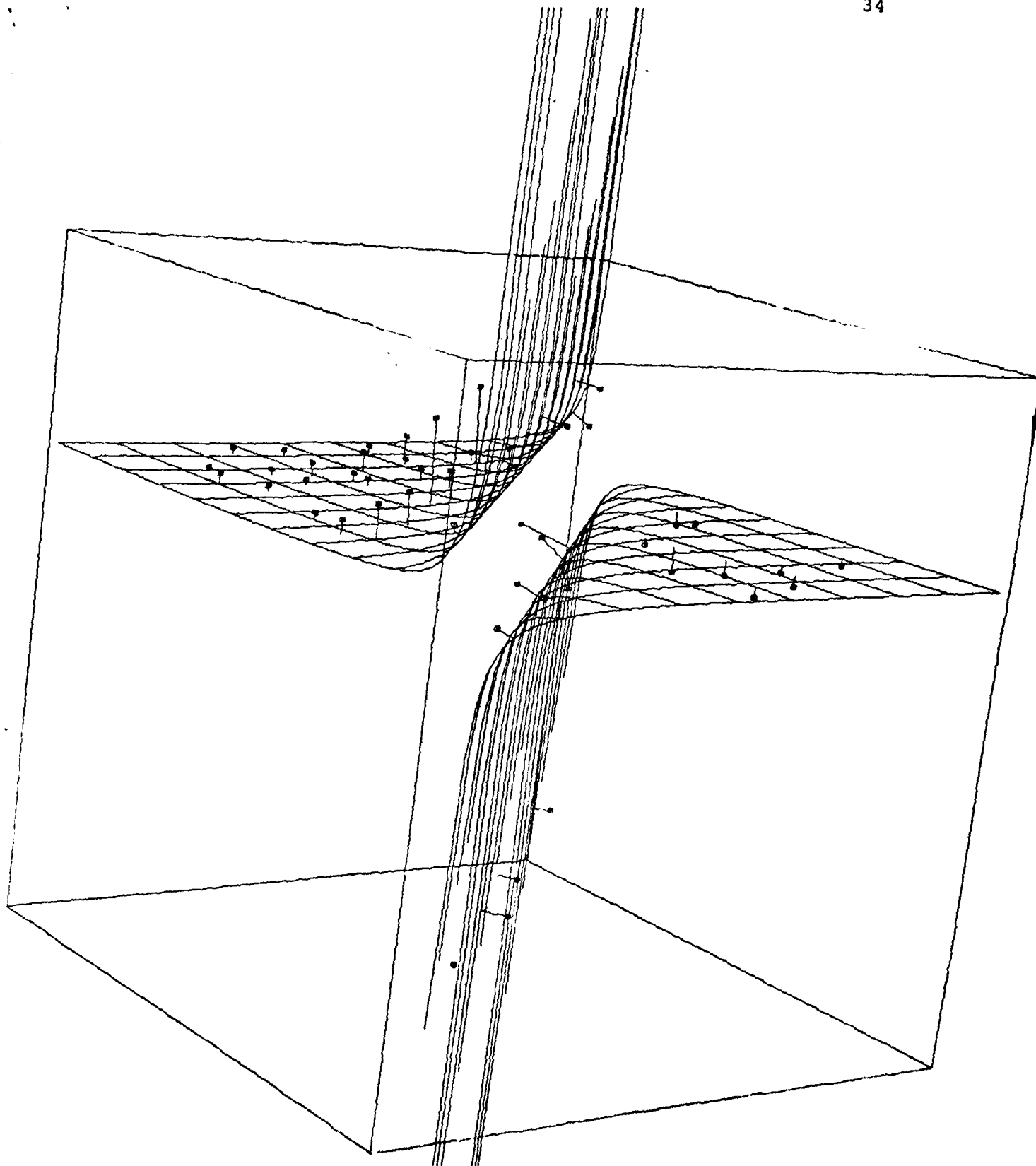


Figure 10

 $\sigma = 100$

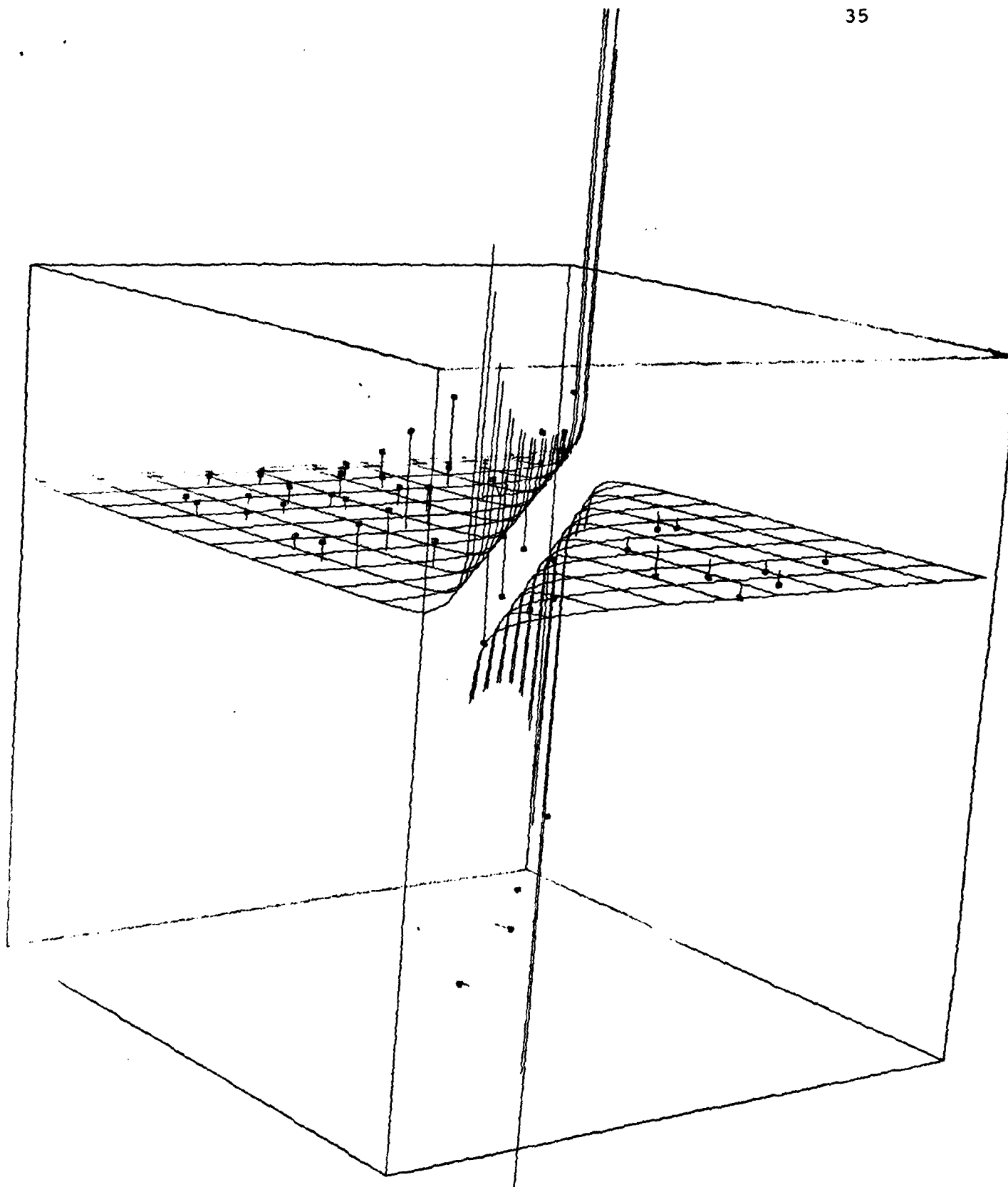


Figure 11
 $d = 500$

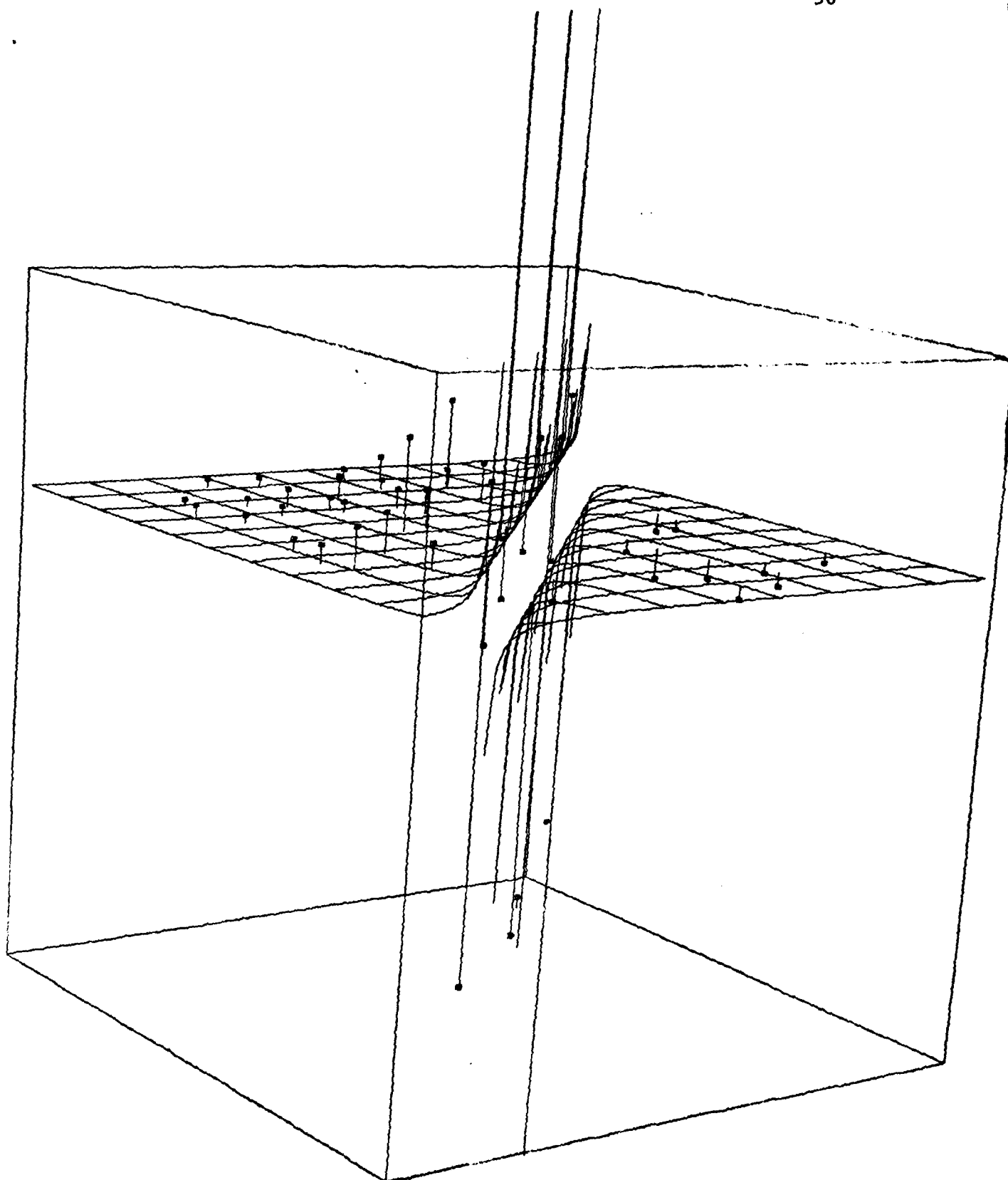


Figure 12
OLS

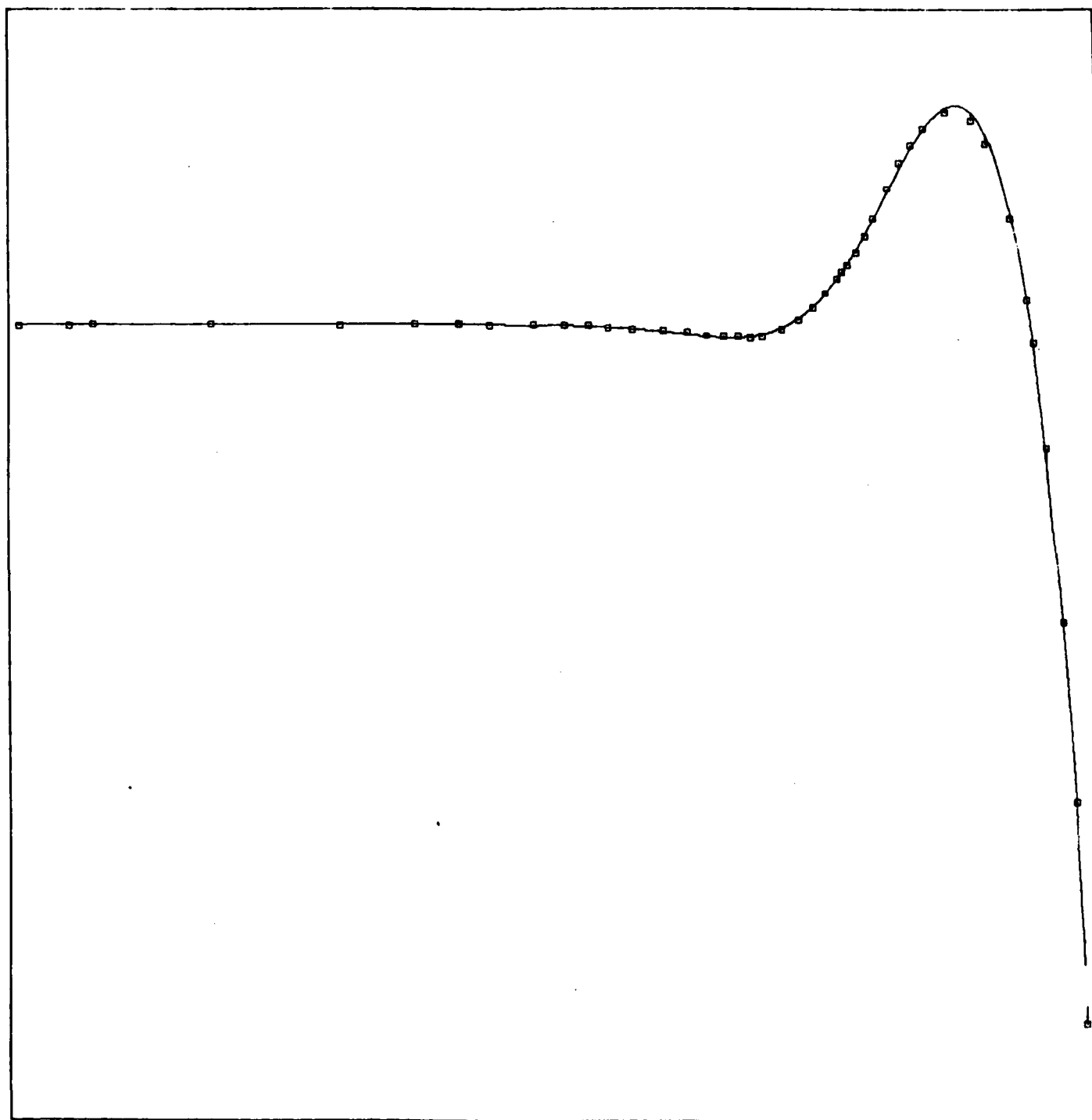


Figure 13
OLS

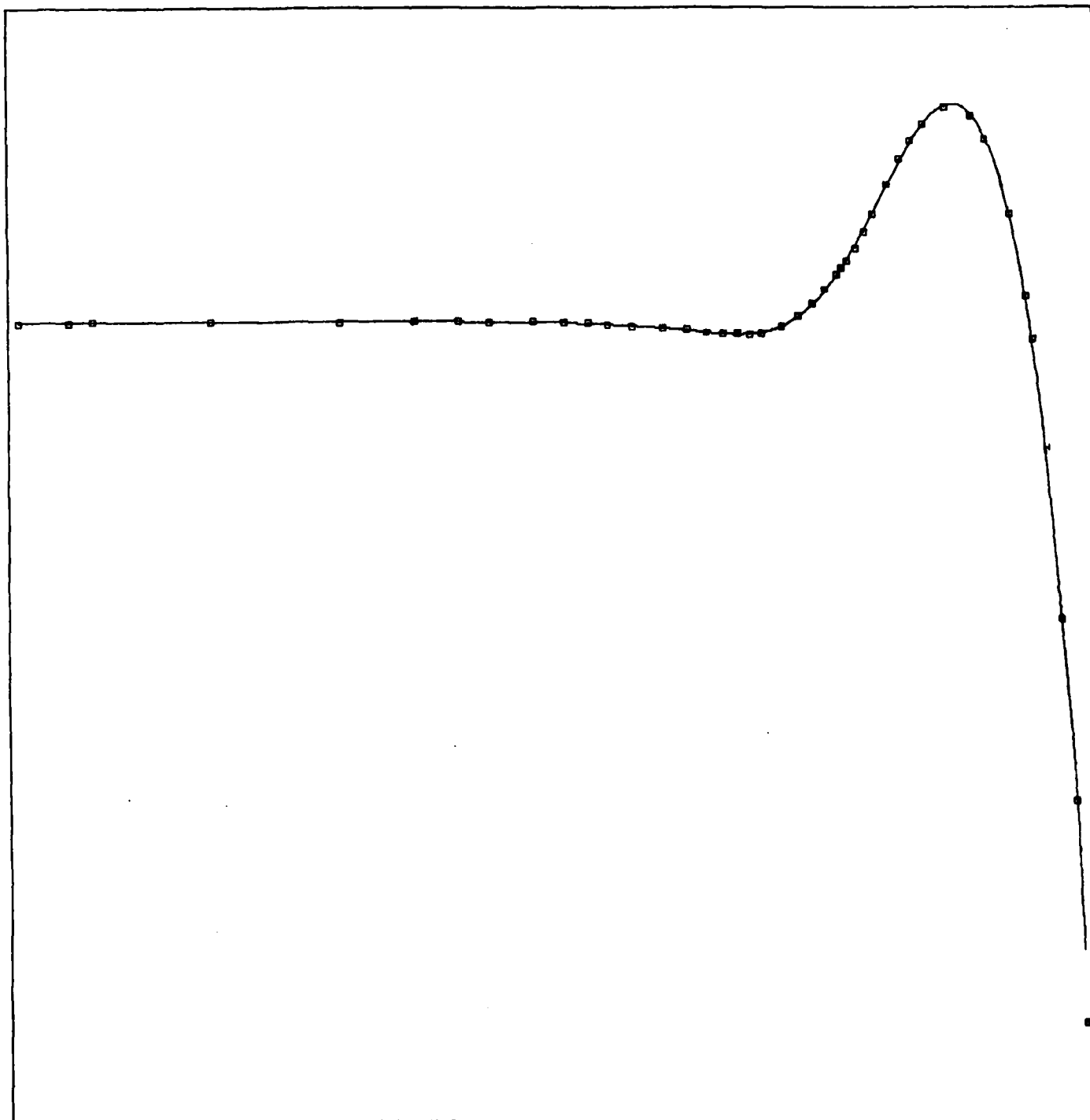


Figure 14
 $\sigma = .01$

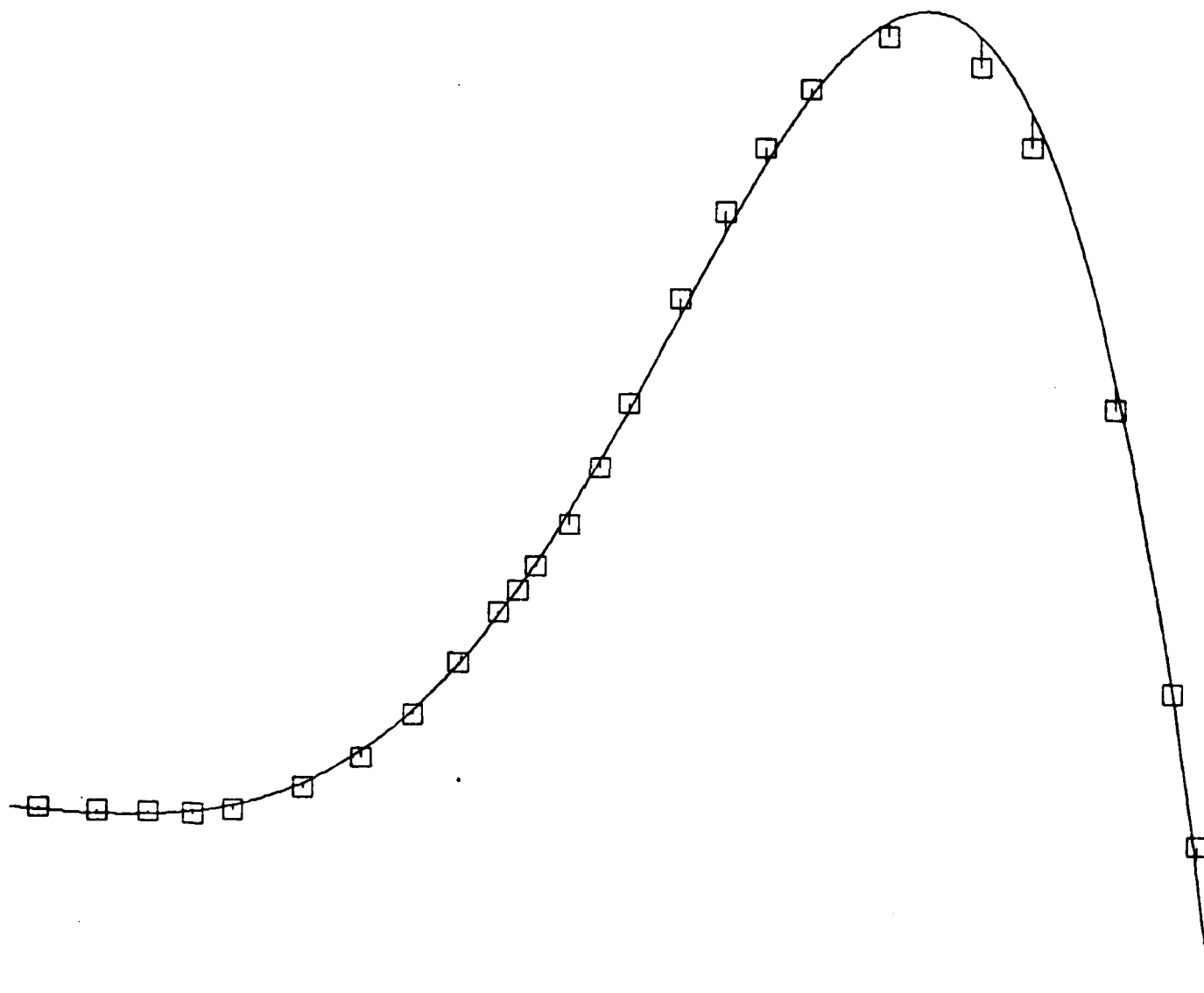


Figure 15
OLS

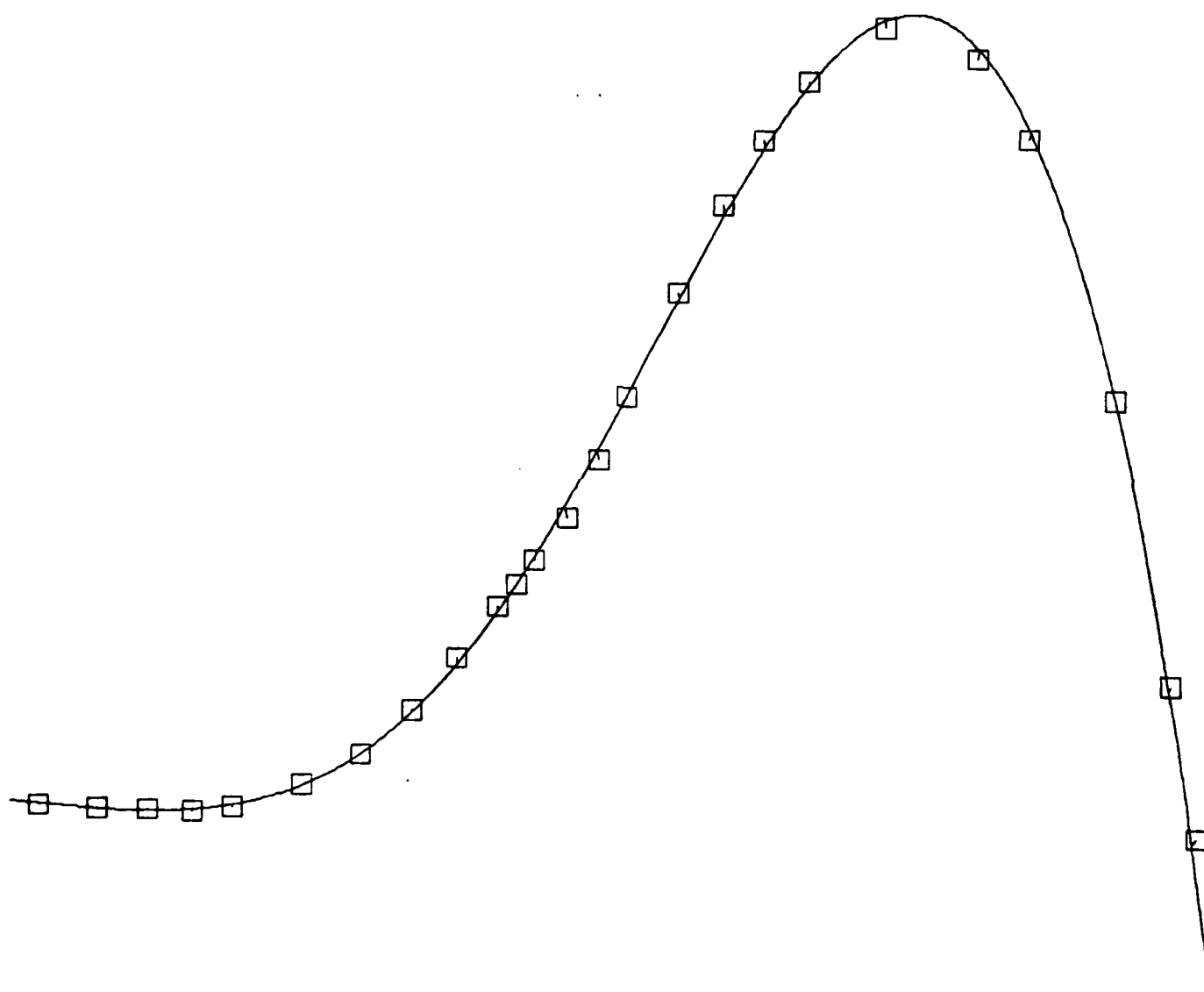


Figure 16
 $\sigma = .01$

END

FILMED

386

DTIC